

DOCTORAL THESIS

2017



**Recent Advances in Ontology-based
Semantic Similarity Measures and
Information Content Models based on
WordNet**

JUAN JOSÉ LASTRA DÍAZ

DOCTORAL PROGRAMME ON INTELLIGENT SYSTEMS

Supervisor: Prof. Dr. Ana García Serrano

Lastra-Díaz, Juan J. (2017). Recent Advances in Ontology-based Semantic Similarity Measures and Information Content Models based on WordNet. Universidad Nacional de Educación a Distancia (UNED). PhD Thesis.

© 2017 by Juan José Lastra Díaz

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.



A Fátima, mi maravillosa,
amada y paciente esposa,
por haber soportado
estoicamente mi dedicación
a la etapa que culmina
con este trabajo.

A mis queridos hijos,
Fátima, Iñigo y Jaime,
a quienes he robado
mucho atención
durante estos años.

Contents

Abstract	vii
Resumen	ix
Acknowledgements	xiii
I Thesis by Compendium	1
1 Introduction	3
1.1 Ontologies versus corpus	5
1.2 Definition of the research problem	6
1.3 Structure of this thesis	6
2 Hypotheses and Objectives	9
2.1 A new family of semantic similarity measures	9
2.1.1 Main motivation and hypothesis	9
2.1.2 Research problem and objectives	10
2.2 A new family of IC models	11
2.2.1 Main motivation and hypothesis	12
2.2.2 Research problem and objectives	12
2.3 A refinement of our family of IC models	13
2.3.1 Main motivation and hypothesis	13
2.3.2 Research problem and objectives	14
2.4 Efficient and scalable reproducibility resources	15
2.4.1 Main motivation and hypothesis	16
2.4.2 Research problems and objectives	17
2.5 Evaluation of our new IR model	18
2.5.1 Main motivation and hypothesis	20
2.5.2 Research problem and objectives	22
2.5.3 Evaluation problems leading to abandoning this task	22
3 Theoretical Foundations and Methodology	25
3.1 Theoretical foundations	25
3.2 Research methodology	27

4	Conclusions and Future Work	29
4.1	Main conclusions	29
4.2	Future work	35
5	Scientific contributions	37
5.1	Peer-reviewed articles	37
5.2	Technical reports	37
5.3	Patent applications	38
5.4	Software libraries	38
5.5	Replication datasets and benchmarks	38
6	Impact factor of the publications	39
II	Publications and Patents	55
7	Engineering Applications of Artificial Intelligence article	57
8	Knowledge-Based systems article	75
9	UNED Technical Report	103
10	Information Systems article	151
11	US Patent Application	175
III	Software Libraries and Datasets	225
12	HESML V1R2 Semantic Measure Library	227
13	HESML V1R1 Semantic Measure Library	231
14	WNSimRep V1 dataset	235
15	Reproducible Experiments based on ReproZip	255
16	Benchmarks between Semantic Measure Libraries	259

List of Tables

1.1	Categorization of the main ontology-based semantic similarity measures based on WordNet, and two other measures based on Wikipedia and Linked Open Data (LOD) respectively.	8
4.1	Results obtained for the main hypotheses and research questions studied by this thesis.	36
6.1	2-year JCR impact factors of the three main publications derived from this thesis. All of the above journals are edited by Elsevier.	39

List of Figures

3.1	Research methodology adopted in this thesis. We place a special emphasis in the replications of previous methods and results, as well as their confirmation and refutation.	28
6.1	JCR-2015 Impact Factor and Quartile of our three main publications (source: WoK-FECYT)	40
6.2	JCR-2016 impact factor of our three main publications (source: WoS InCites)	40

Abstract

Human similarity judgments between concepts underlie most of cognitive capabilities, such as categorization, memory, decision-making and reasoning. Thus, the proposal for concept similarity models to estimate the degree of similarity between word and concept pairs has been a very active line of research with many applications in the fields of cognitive sciences, artificial intelligence, Information Retrieval (IR) and genomics, among others. The most successful approach to estimate human similarity judgements is set by the family of ontology-based semantic similarity measures based on WordNet for general domain applications, or MeSH and SNOMED for biomedical applications, as well as the Gene Ontology (GO) for genomics. The advent of the Semantic Web has encouraged the emergence of a novel family of IR models and semantic search systems based on ontologies. In this latter scenario, the ontologies have also been extensively used as semantic conceptual spaces with the aim of indexing and representing large collections of documents and other types of semantically-annotated information.

This thesis introduces two new families of ontology-based semantic similarity measures and Information Content (IC) models based on WordNet together with the largest experimental surveys reported in the literature. Our experiments are based on our software implementation of most methods reported in the literature. In addition, this thesis introduces several significant contributions into the reproducibility of word similarity benchmarks, ontology-based semantic similarity measures and IC models as follows: (1) a new and efficient representation model for taxonomies, called *PosetHERep*, which is an adaptation of the half-edge data structure commonly used to represent discrete manifolds and planar graphs; (2) a new Java software library called the *Half-Edge Semantic Measures Library (HESML)* based on *PosetHERep*, which implements most ontology-based semantic similarity measures and IC models reported in the literature; (3) a set of reproducible experiments on word similarity based on *HESML* and *ReproZip* with the aim of exactly reproducing the experimental surveys in all our previous works; (4) a replication framework and dataset, called *WNSimRep v1*, whose aim is to assist in the exact replication of most methods reported in the literature; and finally, (5) a set of scalability and performance benchmarks for semantic measure libraries.

Our novel family of ontology-based semantic similarity measures is based on two previously unconsidered notions as follows: a generalization of the classic Jiang-Conrath (J&C) distance to any type of taxonomy which is based on an IC-based weighted graph derived from the conditional probabilities between child and parent concepts, and a non-linear normalization function that converts the ontology-based semantic distances into similarity functions. Likewise, our new family of intrinsic

and corpus-based IC models is based on two previously unconsidered notions as follows: the preservation of the probabilistic structure of the taxonomy associated to the conditional probabilities between child and parent concepts, and the explicit consideration of a cognitive similarity notion in the definition of the IC model.

Our new IC-based similarity measures outperform the state-of-the-art measures in a statistically significant manner, whilst our new family of IC models obtains rivaling results as regards the state-of-the-art methods and sets an open framework for the derivation of novel intrinsic IC models based on alternative methods for the estimation of the conditional probability between child and parent concepts. On the other hand, *PosetHERep* proposes a memory-efficient representation for taxonomies which linearly scales with the size of the taxonomy and provides an efficient implementation of most taxonomy-based algorithms used by the semantic measures and IC models, whilst *HESML* provides an open framework to aid research into the area by providing a simpler and more efficient software architecture than the current software libraries. *HESML* outperforms the state-of-the-art semantic measure libraries by several orders of magnitude and shows that it is possible to improving their performance and scalability significantly without caching using *PosetHERep*. Our large experimental surveys, including most similarity measures and IC models based on WordNet reported in the literature, also led us to be on the lookout for several reproducibility problems in the replication of methods and experiments previously reported in the literature, as well as the discovery of contradictory results. Likewise, our experimental surveys allow us to refute two common beliefs held among the research community: (1) a wrong belief about the outperformance of intrinsic IC models over those based on a corpus that is refuted by our results, and (2) another wrong belief about the overall outperformance of the classic IC-based similarity measures on the family of path-based semantic measures, which is refuted by our conclusion that only a small set of similarity measures based on recent hybrid IC-based measures obtain a statistically significant higher Spearman correlation value than the family of path-based similarity measures. This latter fact explains some unexpected results in information retrieval applications based on similarity measures in which several authors point out that there is no a statistically significant difference between the performance obtained by the families of classic semantic similarity measures based on IC models and other classic measures based on the length of the shortest path between concepts when the Spearman correlation metric is used.

Keywords: ontology-based semantic similarity measures, intrinsic and corpus-based Information Content models, WordNet-based semantic similarity measures, ontology-based IR models, HESML, PosetHERep, semantic measures library, reproducible experiments on word similarity, WNSimRep v1 dataset, ReproZip, replication datasets for ontology-based semantic similarity models

Resumen

Los juicios de semejanza entre conceptos subyacen tras la mayoría de capacidades cognitivas, tales como la categorización, la memoria, la toma de decisiones y el razonamiento. Por lo tanto, la propuesta de modelos de semejanza conceptual para estimar el grado de semejanza entre pares de palabras y conceptos ha sido una línea muy activa de investigación, con muchas aplicaciones en los campos de las ciencias cognitivas, la inteligencia artificial, la recuperación de la información (RI) y la genómica, entre otros. El enfoque de mayor éxito para estimar juicios de semejanza es definido por la familia de medidas de semejanza semántica basadas en ontologías para dominios generales de aplicación basados en WordNet, o MeSH y SNOMED para aplicaciones biomédicas, así como la Ontología Génica (GO) para genómica. El advenimiento de la Web Semántica ha motivado la aparición de una nueva familia de modelos de recuperación de la información y sistemas de búsqueda semántica basados en ontologías. En este último escenario, las ontologías han sido extensivamente utilizadas como espacios conceptuales con el propósito de indexar y representar grandes colecciones de documentos y otros tipos de información anotada semánticamente.

Esta tesis presenta dos nuevas familias de medidas de semejanza semántica basadas en ontologías y modelos de contenido de la información basados en WordNet, junto con los mayores estudios experimentales publicados. Nuestros experimentos se basan en nuestra propia implementación de la mayoría de métodos publicados. Adicionalmente, esta tesis presenta algunas contribuciones significativas en la reproducibilidad de estudios experimentales de semejanza entre palabras, medidas de semejanza semántica basadas en ontologías y modelos de contenido de la información, tales como: (1) un nuevo y eficiente modelo de representación para taxonomías, denominado *PosetHERep*, el cual es una adaptación de la estructura de datos ‘half-edge’, utilizada comunmente para representar variedades discretas y grafos planos; (2) una nueva biblioteca de software en Java, denominada *Half-Edge Semantic Measures Library (HESML)*, basada en *PosetHERep*, la cual implementa la mayoría de medidas de semejanza semántica basadas en ontologías y modelos de contenido de la información reportados en la literatura; (3) un conjunto de experimentos reproducibles de semejanza entre palabras basados en *HESML* y *ReproZip*, con el propósito de reproducir de manera exacta los experimentos publicados en todos nuestros trabajos anteriores; (4) un marco y conjunto de datos de replicación, denominado *WNSimRep v1*, cuyo objetivo es ayudar en la replicación exacta de la mayoría de métodos publicados; y por último, (5) un conjunto de estudios experimentales de rendimiento y escalabilidad para librerías de medidas semánticas.

Nuestra nueva familia de medidas de semejanza basadas en ontologías está ba-

sada en dos nociones no consideradas con anterioridad: una generalización de la distancia clásica de Jiang-Conrath a cualquier tipo de taxonomía, la cual se basa en un grafo pesado basado en un modelo de contenido de la información derivado de las probabilidades condicionales entre conceptos padres e hijos, y una función de normalización no lineal que convierte las medidas de distancia semántica basadas en ontologías en funciones de semejanza. Asimismo, nuestra nueva familia de modelos de contenido de la información de tipo intrínseco y basados en corpus se basa en dos nociones no consideradas previamente: la preservación de la estructura probabilística de la taxonomía asociada a las probabilidades condicionales entre conceptos padre e hijos, y la consideración explícita de una noción de semejanza cognitiva en la definición del modelo de contenido de la información.

Nuestras nuevas medidas de semejanza basadas en modelos de contenido de la información superan de manera estadísticamente significativa a las medidas estado del arte, mientras que nuestra nueva familia de modelos de contenido de la información obtiene resultados comparables con respecto a los métodos estado del arte y define un marco abierto para la derivación de nuevos modelos intrínsecos de contenido de la información basados en métodos alternativos para la estimación de las probabilidades condicionales entre conceptos padre e hijos. Por otra parte, *PosetHERep* propone un modelo eficiente de representación para taxonomías respecto al uso de memoria, el cual escala linealmente con el tamaño de la taxonomía y ofrece una implementación eficiente de la mayoría de algoritmos basados en taxonomías que son empleados por las medidas semánticas y los modelos de contenido de la información, mientras que *HESML* ofrece un marco abierto para ayudar en la investigación en el área ofreciendo una arquitectura de software más sencilla y eficiente que las bibliotecas de software actuales. *HESML* supera a las bibliotecas de medidas semánticas actuales por varios órdenes de magnitud y prueba que es posible mejorar significativamente su rendimiento y escalabilidad sin utilizar almacenamiento auxiliar mediante el uso de *PosetHERep*. Nuestros grandes estudios comparativos, incluyendo la mayoría de medidas de semejanza y modelos de contenido de la información publicados, también nos conducen a alertar sobre algunos problemas de reproducibilidad en la replicación de métodos y experimentos publicados previamente, así como al descubrimiento de resultados contradictorios. Asimismo, nuestros estudios experimentales nos permiten refutar dos creencias comunes mantenidas entre la comunidad científica: (1) una creencia errónea sobre la ventaja de rendimiento de los modelos de contenido de la información de tipo intrínseco sobre los basados en corpus que es refutada por nuestros resultados, y (2) otra creencia errónea sobre la ventaja global de las medidas clásicas de semejanza basadas en modelos de contenido de la información sobre la familia de medidas semánticas basadas en caminos, la cual es refutada por nuestra conclusión de que sólo un pequeño conjunto de medidas híbridas recientes de semejanza basadas en modelos de contenido de la información obtiene una correlación de Spearman de manera estadísticamente significativa mayor que la familia de medidas de semejanza basadas en caminos. Este último hecho explica algunos resultados inesperados en aplicaciones de recuperación de la información basadas en medidas de semejanza en las cuales algunos autores señalan que no existe una diferencia estadísticamente significativa entre el rendimiento obtenido por las familias de medidas de semejanza clásicas basadas en modelos de contenido de la información y otras medidas clásicas basadas en la longitud del camino más corto entre conceptos

cuando se emplea la métrica de correlación de Spearman.

Palabras clave: ontology-based semantic similarity measures, intrinsic and corpus-based Information Content models, WordNet-based semantic similarity measures, ontology-based IR models, HESML, PosetHERep, semantic measures library, reproducible experiments on word similarity, WNSimRep v1 dataset, ReproZip, replication datasets for ontology-based semantic similarity models

Acknowledgements

“Three passions,
simple but overwhelmingly strong,
have governed my life: the longing for love,
the search for knowledge, and unbearable
pity for the suffering of mankind. ”
Bertrand Russell

This part-time PhD thesis is the culmination of a very old dream, a personal challenge that is closely related to my own being. I simply love to do research, and I may not be completely happy doing other thing. It is what I have done for the the last twenty five years in industry, as well as in my previous postgraduate studies and PhD thesis which I abandoned in 1999 to get married. For all that time, I longed to achieve this challenge while waiting for the opportunity to resume my PhD studies. I have enjoyed this learning road a lot, the rigour and formal aspects associated to academic communication and critical thinking being especially significant for me, as well as the rigour of the review processes and the social interaction linked to the scientific exchange.

Firstly, this achievement would have been impossible without the unconditional support of my devoted wife Fátima, and my three children: Fátima, Iñigo and Jaime. I will never be able to make up the stolen time, but I will try. Secondly, I would like to express my most sincere gratitude to my thesis advisor, Prof. Dr. Ana García-Serrano. Since the very beginning, I have benefitted from her wise advice, as well a frank and open communication together with healthy criticism. She helped me a lot to refine my academic writing style, reviewing our manuscripts once and again, and giving me the freedom to make my own decisions in all aspects of my research work. She is an excellent human being with a great emotional intelligence for managing human resources. I am grateful to having shared my learning road with her.

While travelling the road in search of my dream, I have had the fortune of meeting and exchanging impressions with a lot of people in the research community. Most of my exchanges were by email, although from time to time I had the opportunity to hold some exciting face-to-face discussions. Unfortunately, as a part-time PhD student I missed a lot of opportunities to exchange ideas with other members of my research community at the UNED. For this reason, my sporadic academic discussions, even by email, have likely been the most rewarding part of my research and they have greatly influenced my work. My main working references were the contributions made by Ted Pedersen, David Sánchez, Montserrat Batet, Sébastien

Harispe, Mohammed Hadj Taieb, Peter Gärdenfors, Dominic Widdows and Miriam Fernández. Ted Pedersen kindly answered all our questions and provided us with the WordNet-based frequency files used to build all the corpus-based IC models used in our experiments, whilst Sébastien Harispe provided the SML source code and his total support in evaluating it. Jian-Bo Gao, David Sánchez, Montserrat Batet and Giuseppe Pirró kindly answered all our questions to clarify certain issues on their methods and experimental results with the aim of replicating them in our experimentation platform. Mohamed Hadj Taieb kindly offered us his total support in replicating their similarity measures exactly. Miriam Fernández gave us her wise and timely advice on the current evaluation difficulties in the family of ontology-based Information Retrieval (IR) models, which allowed us an early and successful redirection of our original research plan. Jorge Martínez-Gil, Emmanuel Pothos, Emiel Van Miltenburg and Lubomir Stanchev kindly answered our questions about their methods, in addition to holding interesting discussions with me. Rajendra Banjade kindly answered all our questions to clarify the corpus-based IC models used in his experiments. I had a fluid exchange with Abdulgabbar Saif while both were working to complete our PhD theses, which he completed recently with outstanding contributions. Moreover, I had several interesting discussions with Ángel Castellanos who proposed us the use of HESML to manage FCA-based applications and introduced us to Wikidata. Prof. Horacio Rodríguez shared his knowledge on word embedding methods with me. My inquiries on the categorization of concepts led me to contact several researchers into cognitive sciences, that is how I met Cristóbal Calvo, who shared his research into the use of semantic categorization tasks for the early detection of Alzheimer's disease with us, and Prof. Herminia Peraita, who kindly introduced me in the fascinating theory of categorization developed in the field of cognitive psychology, in which she has been involved since the pioneering times of Eleanor Rosch. In addition, Prof. Peraita has given me part of her bibliographic archive on the topic. Fátima Sánchez Cabo invited us to give a talk at the CNIC Institute (CSIC) with the aim of exchanging our results and searching for collaboration opportunities in the field of bioinformatics. Sixto Jansa, as well as the rest of members of the Technology Transfer Office of the UNED (OTRI), gave us their total support to submit our US patent application and registering the source code of our software libraries. Alexis Moreno-Pulido helped us to find many old papers. Mark Hallett carried-out outstanding work checking the proper use of English in all our publications, for this reason, I am very grateful to him for improving my academic writing in English.

I would also like to express my most sincere gratitude to the editors of the Information Systems journal, Dennis Shasha and Fernando Chirigati, for their kind invitation to submit a reproducibility article on our families of ontology-based semantic similarity measures and Information Content (IC) models which allowed us to introduce several significant original contributions for the first time, such as: a new scalable and efficient representation model for taxonomies and a new software library of semantic similarity measures based on the former one, which sets out the new state of the art in terms of performance and scalability in a very conclusive manner. We learned a lot on reproducibility for science, including the finest details which must be considered to submit easily reproducible papers capable of reproducing the methods and experiments exactly. I am a firm believer and sponsor of the

valuable initiative into computational reproducibility introduced by the editors of Information Systems. I firmly think that this initiative should be adopted by all high-quality scientific journals with the aim of setting a new standard in scientific communication. For this reason, I also subscribe to the recent manifesto for a reproducible science introduced by Munafò et al. [99], because good science only can be built on sound foundations.

Finally, I express my most sincere gratitude to all of the anonymous reviewers of our publications for their wise remarks and suggestions with the aim of improving the quality of our articles. Likewise, I also express my most sincere gratitude to those reviewers that rejected any of our submissions, because they were always fair but rigorous and their remarks contributed significantly to improving my academic writing skills, as well as internalizing those features that contribute to writing high-quality papers. I greatly appreciate the immense service provided by the peer-reviewing process to science. For this reason, I willingly accepted four invitations to act as reviewer in which I tried to reciprocate with the same dedication, rigour and wise advice that I always received from my reviewers. Definitively, my reviewers have contributed a lot to my learning process, helping me to model my critical mindset and embracing the rigour and precision which are essential features of the scientific communication, in addition to appreciating the beauty and satisfaction of work well done, I will always be indebted to them.

Juan José Lastra-Díaz
Madrid, 18th June 2017

Part I
Thesis by Compendium

Chapter 1

Introduction

Human similarity judgments between concepts underlie most cognitive capabilities, such as categorization, memory, decision-making and reasoning. Thus, the proposal for concept similarity models to estimate the degree of similarity between word and concept pairs has been a very active line of research in the fields of cognitive sciences [138, 117], artificial intelligence and Information Retrieval (IR) [119]. The semantic similarity measures estimate the degree of similarity between concepts by considering only ‘is-a’ relationships, whilst the semantic relatedness measures also consider any type of co-occurrence relationship. For instance, a *wheel* is closely related to a *car* because the wheels are part of any car; however, a *wheel* is neither a car nor derives from another common close concept such as a *vehicle*, thus their degree of similarity is low. Whilst hand-coded taxonomies, such as WordNet and other sources of knowledge, can be efficiently and reliably used to retrieve the ‘is-a’ relationships between concepts and words, the co-occurrence relationships required by the semantic relatedness measures need to be retrieved from a large corpus.

An ontology-based semantic similarity measure is a binary concept-valued function $sim : C \times C \rightarrow \mathbb{R}$ defined on a single-root taxonomy of concepts (C, \leq_C) , which returns an estimation of the degree of similarity between concepts as perceived by a human being. The ontology-based similarity measures have become both a very active area of research, and a key component in many applications. For instance, in the fields of Natural Language Processing (NLP) and IR, ontology-based semantic similarity measures have been used in Word Sense Disambiguation (WSD) methods [106], text similarity measures [94], spelling error detection [17], sentence similarity models [104, 71, 46], paraphrase detection [39], unified sense disambiguation methods for different types of structured sources of knowledge [82], document clustering [27], ontology alignment [26], document [86] and query anonymization [11], clustering of nominal information [10, 9], chemical entity identification [43], interoperability between agent-based systems [34], and ontology-based Information Retrieval (IR) models such as that proposed by Lastra-Díaz [58] to solve the lack of an intrinsic semantic distance in vector ontology-based IR models [19]. In the field of bioengineering, ontology-based semantic similarity measures have been proposed for synonym recognition [20] and biomedical text mining [12, 110, 127]. However, since the pioneering work of Lord et al. [80], the proposal of similarity measures for genomics and proteomics based on Gene Ontology (GO) [6] have attracted a lot of attention, as detailed in a recent survey into the topic by Mazandu et al. [88]. Many

GO-based semantic similarity measures have been proposed for protein functional similarity by several authors [25, 24, 113, 145], giving rise to its applications in protein classification and protein-protein interactions [142, 44], gene prioritization [131] and many others reported in [88, p.2].

Given a taxonomy of concepts defined by the triplet $\mathcal{C} = ((C, \leq_C), \Gamma)$, where $\Gamma \in C$ is the supreme element called the root, an Information Content model is a function $IC : C \rightarrow \mathbb{R}^+ \cup \{0\}$, which represents an estimation of the information content for every concept, defined by $IC(c_i) = -\log_2(p(c_i))$, $p(c_i)$ being the occurrence probability of each concept $c_i \in C$. Each IC model must satisfy two further properties: (1) nullity in the root, such that $IC(\Gamma) = 0$, and (2) growing monotonicity from the root to the leaf concepts, such that $\forall c_i \leq_C c_j \Rightarrow IC(c_i) \geq IC(c_j)$. Once the IC-based similarity measure is chosen, the IC model is mainly responsible for the definition of the notion of similarity and distance between concepts.

Current ontology-based semantic similarity measures can be categorized into four subfamilies as shown in table 1.1. First, edge-counting measures, the so-called path-based measures, whose core idea is the use of the length of the shortest path between concepts as an estimation of their degree of similarity, such as the pioneering work of Rada et al. [119]. Second, the family of IC-based similarity measures, whose core idea is the use of an Information Content (IC) model, such as the pioneering work of Resnik [121], and the subsequent measures introduced by Jiang and Conrath [53] and Lin [78]. Third, the family of feature-based similarity measures, whose core idea is the use of set-theory operators between the feature sets of the concepts, such as the pioneering work of Tversky [138]. And fourth, other similarity measures that cannot be directly categorized into any previous family, which are based on similarity graphs derived from WordNet [136], novel contributions of the hyponym set [45], or aggregations of other measures [87]. In turn, the more recent IC-based similarity measures can be divided into four subgroups: (1) a first group made up of the aforementioned three classic IC-based similarity measures by Resnik [121], Jiang and Conrath [53] and Lin [78]; (2) a second group defined by those measures that make up an IC model with any function based on the length of the shortest path between concepts, such as the pioneering work of Li et al. [75], and other subsequent works such as [147], [91], [40] and [60]; (3) a third group of IC-based measures based on the reformulation of different approaches, such as the IC-based reformulations of the Tversky measure by Pirró [114], and the IC-based reformulation of most edge-counting methods introduced by Sánchez and Batet [127]; and finally, (4) a fourth group of IC-based measures based on a monotone transformation of any classic IC-based similarity measure, such as the exponential-like scaling of the Lin measure introduced by Meng and Gu [89], the reciprocal similarity measure of the Jiang-Conrath distance introduced by Garla and Brandt [42], another exponential-like normalization of the Jiang-Conrath distance introduced by Lastra-Díaz and García-Serrano [60], and the monotone transformation of the Lin measure called FaITH introduced by Pirró and Euzenat [115].

On the other hand, the ontologies have found one of their most significant applications in the development of semantic search systems for the Semantic Web or any other type of semantically annotated corpus. It has encouraged the proposal of ontology-based IR models, such as the pioneering work introduced by Castells et al. [19], which are closely connected with the ontology-based semantic measures as the

latter can be used as metrics of any conceptual space derived from a base ontology.

Path-based measures	{	Rada et al. [119], Wu and Palmer [143]
		Leacock and Chodorow [70], Hirst and St-Onge [52]
		Pedersen et al. [110], Al-Mubaid and Nguyen [3]
IC-based measures	{	Classic IC-based measures {
		Resnik [121]
		Jiang and Conrath [53]
		Lin [78]
Hybrid (path-based) IC-based measures	{	Li et al. [75]
		Zhou et al. [146]
		Meng et al. [91]
		Gao et al. [40]
Reformulations of other types of measure	{	Pirró [114]
		Sánchez and Batet [127]
Monotone transformations of classic IC-based measures	{	Pirró and Euzenat [115]
		Meng and Gu [89]
		Garla and Brandt [42]
		Lastra-Díaz and García-Serrano [60] (cosJ&C)
Feature-based measures	{	Tversky [138]
		Batet et al. [12]
		Sánchez et al. [130]
Other types of measure	{	- Taxonomical features (hyponym sets): Hadj Taieb et al. [45]
		- Aggregation of different of measures: Martinez-Gil [87]
		- Asymmetrically weighted graphs based on WordNet: Stanchev [136]
		- IC-based reformulation based on LOD: Meymandpour and Davis [93]
		- IC-based reformulation on Wikipedia: Jiang et al. [54]

Table 1.1: Categorization of the main ontology-based semantic similarity measures based on WordNet, and two other measures based on Wikipedia and Linked Open Data (LOD) respectively.

1.1 Ontologies versus corpus

There are currently two main approaches used to estimate human similarity judgments which can be roughly categorized into two families: ontology-based semantic similarity measures and corpus-based ones. Most of corpus-based similarity measures are based on the distributional hypothesis [50], which states that words in similar contexts tend to share similar meanings. Thus, distributional measures usually define the meanings of the word as a function of their context and the type of

co-occurrence relationships that needs to be captured. For example, the contexts of the word could be small n-gram windows, or larger contexts such as sentences, paragraphs or documents. Despite there being different methods to represent the word meanings (contexts), such as sets, vectors, probability distributions, and graph nodes, the most popular representations rely on vector space models (VSM) [137, §2.2]. However, the mainstream of research in the current family of corpus-based semantic similarity measures is the use of word embeddings, such as those introduced by Mikolov et al. [95] and Pennington et al. [111].

The main advantage of the ontology-based similarity measures is that the logic relationships between concepts, especially the “is-a” relationships, are hand-coded within the ontologies. A second advantage of these measures is that they are defined by closed formulas that only require a taxonomy to be evaluated. Therefore, they can be easily implemented, although their computational cost depends on the size of the ontology and the complexity of the algorithms required. In contrast, a serious drawback of the ontology-based measures in open domain applications, like the Web, is their limited lexical coverage, and the cost of creating and updating wide coverage ontologies. On the other hand, the corpus-based measures mainly rely on the distributional hypothesis, and compute the degree of similarity using an indirect approach that relies on the statistical co-occurrence between word contexts. In addition to the “is-a” relationships, co-occurring words can encode other types of semantic relationships. Therefore, the corpus-based measures “can confuse similarity with relatedness” [74, §1]. Moreover, “it is commonly considered that distributional measures can only be used to capture semantic relatedness” [49, §2.5.2], and “they have traditionally performed poorly when compared to WordNet-based measures” [97, p.1] in the similarity assessment task. Another drawback of corpus-based measures is that they are commonly based on a pipeline of NLP and IR algorithms, as well as external services and resources, resulting in a high computational cost and replication complexity. For instance, a recent paper by Fares et al. [36] details a lot of reproducibility problems in the setup and training of current state-of-the-art word embeddings which encourage the development of a public repository of pre-processed corpora and pre-trained vectors with the aim of making the evaluation and comparison of methods easier. In addition, the corpus-based measures exhibit the classic problems related to corpus statistics, such as the difficulty in obtaining a well-balanced corpus for all words and their senses. In summary, ontology-based similarity measures are efficient, robust and easy to implement, whilst corpus-based measures offer a broader lexical coverage at the expense of a higher computational complexity and many reproducibility difficulties.

1.2 Definition of the research problem

This thesis introduces a compendium of my research tackling the problem of proposing and evaluating new state-of-the-art methods, software tools and reproducibility resources in the family of ontology-based semantic similarity measures based on WordNet, as well as the application of the previous methods and resources in the proposal and evaluation of a new ontology-based Information Retrieval (IR) model for semantic search.

The research in this thesis is divided into two main lines of research which are grouped in workpackages as follows. WP1 tackles the following research problems: (1) the proposal and evaluation of a new family of ontology-based semantic similarity measures based on IC models; (2) the proposal, evaluation and refinement of a new family of intrinsic and corpus-based IC models; (3) the proposal and evaluation of an efficient and scalable representation model for taxonomies together with a new software library of semantic measures based on it; (4) the proposal and evaluation of a new replication framework and dataset for the exact replication of ontology-based semantic similarity measures and IC models based on WordNet; (5) the proposal and evaluation of a set of reproducible experiments into word similarity based on ReproZip [22] and our new semantic measures library. And finally, WP2 tackles the problem of proposing and evaluating a new ontology-based Information Retrieval (IR) model based on our family of *Intrinsic Ontology Spaces* introduced by Lastra-Díaz [58].

Our two main lines of research are closely related because the ontology-based semantic similarity and distance measures can be used as the metrics for any conceptual space derived from a base ontology, which follows that they are amenable to being used in the definition of semantic aware IR models. For instance, Rada et al. [119] introduce the first path-based semantic distance based on ontologies in their pioneering work, as well as an ontology-based IR model for the biomedical domain which uses an ontology-based semantic distance for the first time to compare the concept-based representation of a user query with the representation of the indexed documents. Thus, almost three decades ago, the seminal work of Rada et al. [119] highlighted the close connection between ontology-based semantic distance and semantic information retrieval models. Surprisingly, the pioneering ideas of Rada et al. [119] linking IR models and semantic distances have gone unnoticed for almost three decades until I rediscovered them in my previous MSc thesis [58]. The works of Rada et al. [119] and Tversky [138] are broadly known and cited as the pioneering work of the subfamilies of path-based and feature-based semantic similarity measures based on ontologies; however, the research community in the field of IR overlooked the aforementioned connection between ontology-based semantic similarity measures and IR models.

1.3 Structure of this thesis

This thesis is structured in three parts as follows. Part I is the main body of this thesis by compendium, whilst part II introduces the full-text of all of the publications derived from this thesis, and finally part III introduces our software libraries and datasets. In addition to the Mendeley datasets, this latter part also includes the full-text of a report detailing our WNSimRep v1 replication dataset. In turn, part I is structured as follows. Chapter 2 introduces a summary of the main motivation, hypotheses, research problems and objectives tackled by each publication derived from this thesis. Chapter 3 details the theoretical foundations and domain knowledge of this thesis together with our research methodology. Chapter 4 introduces our main conclusions and forthcoming activities. Finally, chapter 5 enumerates our scientific contributions, both research articles and software libraries and datasets,

whilst chapter 6 shows a summary table detailing the quality metrics of our main publications.

Chapter 2

Hypotheses and Objectives

This chapter introduces a summary of the main motivation, hypotheses, research problems and objectives tackled by each publication derived from this thesis. The chapter is structured in five different sections, each one matching our main publications. In turn, each section is structured in two other subsections in which we detail our main motivation and hypotheses, as well as the research problems and objectives tackled in each publication.

The research tasks of this thesis have been divided into two lines of research, called workpackages, herein WP1 and WP2. The aim of our first workpackage (WP1) is the proposal and evaluation of novel ontology-based semantic similarity measures and information content (IC) models, together with software libraries, resources and datasets for their replication. On the other hand, WP2 focuses the proposal and evaluation of novel ontology-based information retrieval (IR) models. WP1 research is detailed in sections 2.1 to 2.4, whilst WP2 research is detailed in section 2.5.

2.1 A new family of semantic similarity measures

This section introduces a summary of our research carried-out with the aim of proposing a new family of ontology-based semantic similarity measures based on Information Content (IC) models, together with a large experimental survey based on our software implementation of most previous methods reported in the literature.

The research detailed herein matches the content introduced by Lastra-Díaz and García-Serrano [60]. However, one of the new IC-based semantic distances introduced in [60], called *weighed Jiang-Conratgh distance*, is also disclosed as part of our patent application [65, par. 0001]. In addition, the software implementation of all methods replicated and evaluated in our experiments has been made publicly available as part of our novel HESML software library introduced by Lastra-Díaz et al. [68], as well as a set of reproducible experiments provided as supplementary material which allows all of the experiments in our aforementioned paper to be reproduced exactly.

2.1.1 Main motivation and hypothesis

The main motivation for this research is the identification of two drawbacks in the semantic distance introduced by Jiang and Conrath [53]. Firstly, Jiang and Conrath show in their aforementioned paper that their classic semantic distance is equivalent to the length of the shortest path between concepts on a weighted graph derived from the taxonomy, in which the edge weights are set to the IC values of the conditional probabilities between child and parent concepts. However, this relationship has not been explored before in the definition of any ontology-based semantic similarity measure, whilst Orum and Joslyn [105] show that the classic Jiang-Conrath (J&C) semantic distance is only a metric on tree-like taxonomies, thus rebutting the common belief and the original statements made by their authors. Second, we observe an underlying assumption in the literature as regards the conversion of any ontology-based semantic distance, such as the J&C distance, into a similarity measure. In most cases, the J&C distance is converted into a similarity function through a linear mapping, despite this relationship being unknown and probably non-linear. Thus, our two main hypotheses detailed below are as follows:

Hypothesis 1 (weighted Jiang-Conrath distance) *A new semantic distance defined as the length of the shortest path between concepts in a weighted taxonomy whose edge weights are set to the IC value of the conditional probability between its child and parent concepts, or the difference in absolute value of their IC values, could improve the estimation of the human similarity judgements between words and concepts obtained by the current state-of-the-art similarity measures.*

Hypothesis 2 (non-linear normalization) *A new semantic similarity measure defined as a proper non-linear normalization of the classic Jiang-Conrath distance, or our new weighted Jiang-Conrath distance, could improve the estimation of the human similarity judgements between words and concepts obtained by their non-normalized versions.*

2.1.2 Research problem and objectives

In order to bridge the gap detailed in the previous section and evaluating our two main hypotheses 1 and 2, this research tackles the problem of designing a new family of ontology-based semantic similarity measures based on a non-linear normalization and the generalization of the classic Jiang-Conrath distance to non tree-like taxonomies. Our novel family of similarity measures is based on the definition for the first time of an IC-based weighted taxonomy. The main objectives of the research detailed herein are as follows:

1. To propose a novel family of ontology-based semantic similarity measures based on the Information Content theory.
2. To reconsider some previous conclusions on the outperformance of the intrinsic IC model over the corpus-based ones, which rely on the results reported by Patwardhan and Pedersen [107], Pedersen [109] and Pirró [114]. These latter works are the primary sources that prove that the state-of-the-art intrinsic

IC models outperform the corpus-based ones in the WordNet-based similarity tasks. The comparison and conclusions in other subsequent works, such as Sánchez et al. [130] and Yuan et al. [144], rely on these primary sources.

3. To replicate the methods and state-of-the-art results introduced by Gao et al. [40], which question the conclusions of the research community on the outperformance of the intrinsic IC models over the corpus-based ones.
4. To carry-out a large and up-to-date experimental survey for most of the similarity measures on WordNet, which are based on our own software implementation of most IC models and similarity measures reported in the literature.
5. To replicate most previous methods reported in the literature with the aim of confirming or rebutting their results.
6. To check the reproducibility of previous methods reported in the literature, as well as warning on the irreproducibility of others.
7. To evaluate a family of corpus-based IC models derived from some unexplored WordNet-based frequency files included in the Pedersen [108] dataset.
8. To carry-out an experimental study on the impact of the WordNet version on the performance of the similarity measures.
9. To study the performance of the similarity measures based on WordNet on two versions of the RG65 dataset, the classic one Rubenstein and Goodenough [125] and the recent replication carried-out by Pirró [114].
10. To carry-out the most complete and largest experimental survey of intrinsic and corpus-based IC-based similarity measures based on WordNet with the aim of providing a broad view of the state of the art of the problem.

2.2 A new family of IC models

This section introduces a summary of our research carried-out with the aim of proposing a new family of intrinsic and corpus-based IC models for the estimation of human similarity judgements between word and concepts, which is based on the preservation of the probabilistic structure of the taxonomy, encoded by the conditional probability between child and parent concepts.

The research detailed in this section matches the content introduced by Lastra-Díaz and García-Serrano [59]. However, the core ideas of our family of intrinsic IC models and three intrinsic IC models called *CondProbUniform*, *CondProbHypo* and *CondProbLeaves* are also disclosed as part of our patent application [65, par. 0001], in which the three latter IC models are called *IC-JointProbUniform*, *IC-JointProbHypo* and *IC-JointPrbLeaves*. Like the research introduced in section 2.1, the software implementation of all IC models and similarity measures replicated and evaluated in our experiments has been made publicly available as part of our novel HESML software library introduced by Lastra-Díaz et al. [68], as well as a set of reproducible experiments provided as supplementary material which allows all of experiments in our aforementioned paper to be reproduced exactly.

2.2.1 Main motivation and hypothesis

The main motivation for this research is our observation that the conditional probability functions encode some structure axioms that should be satisfied by any intrinsic or corpus-based IC model, but the IC models in the literature do not consider them, with the exception of the work of Sebti and Barfroush [133].

A second motivation for our research is a first attempt at integrating some ideas in cognitive psychology into the IC models. Gärdenfors [41, section 2.8] introduces a conceptual space model based on a Voronoi diagram, with the aim of explaining a number of plausible production mechanisms for the vagueness of concepts and their categorical perception. However, Gärdenfors does not provide a specific metric for this space and the whole family of intrinsic IC models, which is precisely the aim of this paper. In addition, Gärdenfors [41, p. 46] points out that the mechanisms that explain the vagueness notion, also explain another phenomenon in the cognitive perception of categories which can be defined as follows: *the instance of a concept is more quickly perceived as belonging to another category, when the distance from the prototype of the category increases*. We argue that this latter idea can be formulated through the definition of the *cognitive similarity function as a non-linear function of sigmoid type* over the underlying metric of the conceptual space.

Our two main hypotheses detailed below follow directly from our two main aforementioned motivations.

Hypothesis 3 (well-founded IC models) *A new family of intrinsic and corpus-based IC models based on the explicit encoding of the structure axioms derived from the conditional probabilities could improve the performance obtained by current state-of-the-art IC models in semantic similarity tasks, moreover to provide a better understanding of the problem.*

Hypothesis 4 (cognitive conceptual distance) *A new family of intrinsic and corpus-based IC models based on the explicit encoding of a non-linear cognitive distance between parent and child concepts could improve the performance obtained by current state-of-the-art IC models in semantic similarity tasks.*

2.2.2 Research problem and objectives

In order to bridge the gap detailed in previous section and evaluating our two main hypotheses 3 and 4, this research tackles the problem of designing a new family of IC models for ontology-based semantic similarity measures which is based on the explicit encoding of the structure axioms derived from the conditional probabilities by design, as well as sharing a common computational and algebraic structure. The main objectives of the research detailed herein are as follows:

1. To introduce a new family of intrinsic and corpus-based IC models based on the explicit encoding of the structure axioms derived from the conditional probabilities with the aim of preserving the probabilistic structure of the taxonomy.
2. To propose a general computational and algebraic framework for the design and derivation of new intrinsic IC models based on different methods for the estimation of the conditional probability between child and parent concepts.

3. To propose and evaluate a method to integrate a cognitive similarity notion within the IC model based on the notion of non-linear cognitive distance between the concepts detailed above.
4. To carry-out a large experimental survey of IC models and IC-based semantic similarity measures based on WordNet 3.0 and our own software implementation with the aim of replicating most methods and results previously reported in the literature, including the five most significant datasets.
5. To introduce a new comparison between intrinsic and corpus-based IC models with the aim of confirming some previous conclusions on the outperformance of the intrinsic IC models over the corpus-based obtained by Lastra-Díaz and García-Serrano [60].
6. To propose a new baseline for the evaluation of novel intrinsic IC models based on two corpus-based IC models derived from an unexplored WordNet-based frequency file.

2.3 A refinement of our family of IC models

This section introduces a summary of our research carried-out with the aim of solving two drawbacks found in our recent family of well-founded of IC models detailed by Lastra-Díaz and García-Serrano [59], which matches the content introduced by Lastra-Díaz and García-Serrano [61].

The research detailed herein also includes the largest and most conclusive experimental survey into ontology-based semantic similarity measures and IC models reported in the literature. Like the research introduced by Lastra-Díaz and García-Serrano [60] and Lastra-Díaz and García-Serrano [59], the software implementation of all IC models and similarity measures replicated and evaluated in our experiments has been made publicly available as part of our novel HESML software library introduced by Lastra-Díaz et al. [68], as well as a set of reproducible experiments provided as supplementary material which allows all of the experiments in our aforementioned paper to be reproduced exactly.

2.3.1 Main motivation and hypothesis

Our first motivation is the finding of two drawbacks in the main computation algorithm of our family of *well-founded IC models* introduced by Lastra-Díaz and García-Serrano [59]. First, the two intrinsic and cognitive IC models called *CondProbLogistic* and *CondProbCosine* do not satisfy the axiom that constrains the sum of probabilities on the leaf nodes to be 1. It is a consequence of the non-linear transformations applied to the conditional probabilities of these two models, a fact that was already pointed out in our aforementioned work. Second, in some unlikely cases, the ontologies with multiple inheritance could prevent the IC model from satisfying the *growing monotonicity axiom* in concepts with multiple parents. This latest fact means that for some concept pairs $c_i, c_j \in C$, the constraint $c_i \leq_C c_j \Rightarrow IC(c_i) \geq IC(c_j)$ could be violated. In appendix B of our aforementioned work, which is provided as supplementary material, we show that the main

recovery algorithm of our family of well-founded IC models is a sufficient condition for the sum of probabilities over the leaf nodes to be 1, what follows the underlying probability space is well-defined. However, if the taxonomy exhibits multiple inheritance, the probabilities $p(c_i)$ computed by the aforementioned *algorithm 1* could be higher than the probability of any direct parent in some nodes with multiple parents, thus, leading to a violation of the aforementioned growing monotonicity axiom. Thus, our main hypothesis is as follows:

Hypothesis 5 (Fixing of two structural inconsistencies) *The solution to the two aforementioned structural drawbacks of the main computation algorithm of our family of well-founded IC models could lead us to an improvement in their performance, in addition to fixing an algebraic inconsistency that moves the family of well-founded IC models away from their original design principles.*

The second motivation for the research introduced by Lastra-Díaz and García-Serrano [61] is the lack of an updated and exhaustive evaluation of ontology-based similarity measures and IC models in WordNet, as well as the lack of an exhaustive pairwise statistical significance analysis between them. In the literature, we find some out-of-date similarity benchmarks such as that reported by Budanitsky and Hirst [16] and Budanitsky and Hirst [17], and others, more recent but not as exhaustive, such as Hadj Taieb et al. [45]. The largest and most recent word similarity benchmarks based on WordNet are introduced by Lastra-Díaz and García-Serrano [59] and Lastra-Díaz and García-Serrano [60]. However, not all of the hybrid IC-based similarity measures evaluated in the latest work have been previously evaluated with many IC models considered herein and the datasets introduced by Miller and Charles [96], Agirre et al. [2] and Hill et al. [51]. In addition, most ontology-based semantic similarity measures have never been compared through a statistical significance analysis.

Finally, our final motivation is the replication of previous methods and experiments. Most works introducing similarity measures or IC models during the last decade have only implemented or evaluated classic IC-based similarity measures, such as the Resnik, Lin and Jiang-Conrath measures, avoiding the replication of IC models and similarity measures introduced by other researchers. Some works have not included all the details of their methods, or the experimental setup to obtain the published results, thus, preventing their reproducibility. Most works have copied the results published by others. This latest fact has prevented the valuable confirmation of previous methods and results reported in the literature, which is an essential feature of science. This replication problem is especially significant in the current state of the problem, in which there is no convincing winner within the family of intrinsic IC-based similarity measures and the performance margin is very narrow, as concluded in our aforementioned works [59] and [60].

2.3.2 Research problem and objectives

In order to bridge the gap detailed in the previous section and evaluating our main hypothesis 5, as well as providing a conclusive image of the current state of the problem, in the light of the results reported by Lastra-Díaz and García-Serrano

[60] and Lastra-Díaz and García-Serrano [59], this research tackles the problem of designing and evaluating a refinement of the main algorithm of our family of well-founded IC models and providing a larger evaluation of IC models and ontology-based similarity measures than those available in the literature. Thus, the main objectives of the research detailed herein are as follows:

1. To introduce a refinement of our family of well-founded IC models introduced by Lastra-Díaz and García-Serrano [59] with the aim of fixing the two structural drawbacks stated in our previous section.
2. To introduce the largest and most complete evaluation of IC models and ontology-based semantic similarity measures based on WordNet, which will be based on our software implementation of all methods evaluated in our experiments. The survey will include the most recently available datasets on word similarity based on WordNet, as well as a detailed statistical significance analysis using the Pearson and Spearman correlation metrics.
3. To replicate most ontology-based semantic similarity measures and IC models based on WordNet from the pioneering works of Rada et al. [119] and Seco et al. [134].
4. To provide a new and more conclusive image of the current state of the art in the family of ontology-based semantic similarity measures and IC models based on WordNet.

2.4 Efficient and scalable reproducibility resources

This section introduces a summary of our research into new software tools, resources and datasets for the reproducibility of methods and experiments in the family of ontology-based semantic similarity measures and IC models based on WordNet and their applications. The research detailed herein matches the content introduced by Lastra-Díaz et al. [68] and five Mendeley datasets provided as supplementary material as follows:

1. Our main article on the topic [68], which introduces our new representation model for taxonomies, called *PosetHERep*, and our *HESML* software library of ontology-based semantic similarity measures and IC models, are among other significant contributions for the reproducibility in the area.
2. Two different versions (V1R1 and V1R2) of our *HESML* software library which are publicly available at [62] and [63].
3. A set of reproducible experiments of word similarity based on ReproZip [22] available at [67].
4. A replication framework and dataset to reproduce ontology-based semantic similarity measures and IC models, called *WNSimRep v1*, which is available at [66].

5. A set of scalability and performance benchmarks among the state-of-the-art semantic measures libraries available at [64].

Our work introduced by Lastra-Díaz et al. [68] is a reproducible paper which provides a detailed protocol together with the full set of resources detailed above with the aim of allowing the exact replication of all methods and results introduced in our series of works on ontology-base semantic similarity measures and IC models. The reproducibility of the aforementioned paper is certified by three independent reviewers listed as co-authors.

2.4.1 Main motivation and hypothesis

The two main motivations for our research into the reproducibility of methods and experiments into ontology-based semantic similarity measures and IC models are the three drawbacks in the current semantic measures libraries detailed in paragraph below, and the lack of a set of self-contained and easily reproducible experiments into ontology-based semantic similarity measures and IC models based on WordNet. Another significant motivation, also related to the reproducibility, is the lack of a gold standard to assist in the exact replication of ontology-based semantic similarity measures and IC models.

Our first motivation is the discovery of several scalability and performance drawbacks in the current state-of-the-art semantic measures libraries. We argue that these aforementioned drawbacks are derived from the use of naive graph representation models which do not capture the intrinsic structure of the taxonomies being represented. As a consequence of this latter fact, all topological algorithms based on naive representation models demand a high computational cost which degrades their performance. In turn, in order to solve the performance problem of their graph-based algorithms, the current semantic measures libraries adopt a caching strategy, storing the ancestor and descendant sets of all vertices within the taxonomy, among other topological queries in memory or relational databases. This latter caching strategy significantly increases the memory usage and leads to a scalability problem as regards the size of the taxonomy, in addition to impacting the performance because of the further memory allocation and dynamic resizing of the caching data structures, or the interrogation of external relational databases. A second motivation is related to several software architecture issues that lead to practical difficulties for the functional extension of current software libraries, whilst a third motivation is the lack of software implementations for the most recent ontology-based similarity measures and intrinsic IC models developed during the last decade. This latter fact prevents the publication of exhaustive experimental surveys comparing the new proposed methods with most recent methods reported in the literature, because of the effort and difficulty in replicating previous methods and experiments. Thus, the three aforementioned drawbacks lead us to present our main hypothesis and research questions as follows:

Hypothesis 6 (Intrinsic representation model for taxonomies) *A new representation model for taxonomies which properly encodes their intrinsic structure, together with a new software library based on it, should bridge the aforementioned*

gap of scalability and performance of the current state-of-the-art semantic measures libraries.

Research question 1 *Is a new intrinsic representation model for taxonomies able to improve significantly the scalability and performance of the current state-of-the-art semantic measures libraries?*

Research question 2 *Is it possible to improve significantly the scalability and performance of the state-of-the-art semantic measures libraries without using any caching strategy?*

A fourth motivation of this research is the lack of a set of self-contained and easily reproducible experiments that allow the research community to be able to replicate methods and results reported in the literature exactly, even without the need for software coding. The lack of reproducible experiments, together with the aforementioned lack of software libraries covering the most recent methods, and the difficulties in replicating methods and experiments exactly have contributed, with few exceptions, to improveable reproducibility practices in the area.

And finally, our final motivation is the lack of a gold standard to assist in the exact replication of ontology-based similarity measures and IC models. Most ontology-based similarity measures and intrinsic IC models require the computation of different taxonomical features, such as node depths, hyponym sets, node subsumers, the Least Common Subsumer (LCS), and subsumed leaves, among others. WordNet is a taxonomy with multiple inheritance, thus, some of these features are ambiguously defined, or their computation could be prone to errors. For example, the node depth can be defined as the shortest ascending path length from the node to the root, or the longest ascending path length as defined by Hadj Taieb et al. [45, eq. 40, p. 251]. Different definitions of depth also lead us to different values for the LCS concepts. On the other hand, the computation of the hyponym set, subsumed leaves and subsumer set requires a careful counting process to avoid node repetitions, as has already been pointed by Seco et al. [134, section 3]. Another potential source of error is the ambiguity in the definition and notation of some IC models and similarity measures. For example, Zhou et al. [147, table 1, p. 258] define the root depth as 1, whilst the standard convention in graph theory is 0. Most authors define the hyponym set as the descendant node set without including the base node itself. However, in [45], the hyponym set also includes the base concept. In addition, we find works that do not detail the IC models used in their experiments, or how these IC models were built [40, section 4]. Finally, many recent hybrid-type measures also require the computation of the length of the shortest path between concepts. These sources of ambiguity and difficulty demand a lot of attention to the fine details for replicating most IC models and similarity measures in the literature. In a recent work [60], we find some contradictory results and difficulties in replicating previous methods and experiments reported in the literature. These reproducibility problems were confirmed in another subsequent work, such as [59], whilst new contradictory results are reported by Lastra-Díaz and García-Serrano [61, section 6.10]. Several replication problems were solved with the kind support of most authors. However, we were not able to confirm all previous results, whilst others could not be reproduced through lack of information. As we have explained above, many taxonomical

features are ambiguously defined or prone to errors. Thus, all the aforementioned facts lead us to conclude that the exact replication of ontology-based similarity measures and IC models is a hard task, and not exempt from risk. Therefore, it follows that it is urgent and desirable to set of a gold standard for this taxonomical information in order to support the exact replication of the methods reported in the literature.

2.4.2 Research problems and objectives

In order to bridge the gap detailed in the previous section as well as evaluating our main hypothesis 6 and answering our two research questions 1 and 2, this research tackles the problem of designing a scalable and efficient representation model for taxonomies and a new semantic measures library based on the former, as well as the lack of self-contained reproducible experiments on WordNet-based similarity, tools and resources to assist in the exact replication of methods and experiments previously reported in the literature. Thus, the main objectives of the research detailed herein are as follows:

1. To propose and evaluate a new scalable and efficient representation model for taxonomies, called *PosetHERep*, which is an adaptation of the half-edge data structure commonly used to represent discrete manifolds and planar graphs in computational geometry.
2. To propose and evaluate a new Java software library called *Half-Edge Semantic Measures Library (HESML)* based on *PosetHERep*, which implements most ontology-based semantic similarity measures and Information Content (IC) models reported in the literature.
3. To introduce a set of reproducible experiments on word similarity based on *HESML* and *ReproZip* for the first time with the aim of reproducing the experimental surveys reported in [60, 59, 61] exactly, which are provided as supplementary material at [67]. *ReproZip* is a virtualization tool introduced by Chirigati et al. [22], whose aim is to warrant the exact replication of experimental results onto a different system from that originally used in their creation. *ReproZip* captures all the program dependencies and is able to reproduce the packaged experiments on any host platform, regardless of the hardware and software configuration used in their creation. Thus, *ReproZip* warrants the reproduction of the experiments introduced herein in the long term.
4. To propose a replication framework and dataset called *WNSimRep v1*, for the first time, which is provided as supplementary material at [66], and whose aim is to assist in the exact replication of most methods reported in the literature.
5. And finally, to propose a set of scalability and performance benchmarks, for the first time, to evaluate and compare the current state-of-the-art semantic measures libraries available at [64].

2.5 Evaluation of our new IR model

In my previous MSc thesis [58], I propose a novel structure-preserving ontology-based IR model, called *Intrinsic Ontological Spaces*, with the aim of solving several drawbacks of the current family of ontology-based IR models detailed in section 2.5.1. My aforementioned thesis was defended privately with the aim of submitting our patent application [65] before it was made publicly available. However, we submit our patent application without evaluating our aforementioned IR model. For this reason, we set it as the main aim of our second line of research.

This section introduces a summary of our research carried-out with the aim of proposing and evaluating a new ontology-based IR model based on the preservation of a set of intrinsic semantic and geometric structures implicitly encoded by any base ontology used as indexing semantic space. The research introduced in this section matches the content of our patent application [65], which is based on our preliminary ideas in [58]. However, this latter patent application also discloses and protects part of our new family of semantic similarity measures introduced in [60], as well as the core ideas behind our new family of well-founded intrinsic and corpus-based IC models introduced in [59].

The classic Vector Space Model (VSM) introduced by Salton et al. [126] is known as "bag of words", because every document is represented by a vector whose coordinates are defined as a function of the term occurrence frequency within a document. The set of terms used to represent every document is called the vocabulary of the model, and it defines the base vectors of the model. In most of cases, the cosine function is used as a similarity measure between a query vector and the vectors representing the indexed documents. Because of its simplicity, the VSM model has been adopted in many natural language processing (NLP) applications, such as: information retrieval (IR), document categorization (TC) and clustering, web mining and automatic text summarization (TS) among others.

Despite the vector space models having mainly been used to represent text documents, these models have been extended and successfully applied to represent other types of information units, such as words, phrases and sentences, as is reflected in several reviews by Erk [33], Clark [23] and Turney and Pantel [137]. A word or phrase space is a vector space where the vectors represent these information units instead of documents, and the space metric encodes the semantic similarity between information unit pairs. The word spaces are based on the distributional hypothesis [7], which sets that words in similar contexts have similar meanings. In these models, the vectors representing every word are built as a function of the term's frequency in the context of one word within a document, so that these models allow some semantic relationships and statistics to be encoded, such as the term co-occurrence, the synonymy and the meronymy among others.

The main drawback of the classic VSM model is its lack of meaning, as pointed out by Metzler [92, p. 3]. Most current academic information retrieval models use a standard "bag of words" VSM model with meaningless terms, which prevents the retrieval of documents using queries based on non-explicit terms mentioned in the corpus. On the other hand, the same situation occurs in other related problems where the same meaningless version of the VSM model is used, for instance, in text categorization as noted by Sebastiani [132] and Lewis et al. [73].

The advent of the semantic web has encouraged the following change of paradigm among the IR research community: the IR models have moved from a model based on meaningless terms to a model based on references to concepts or their instances. This new paradigm has converted the conceptual models and the knowledge bases in their core components, thus, ontology languages such as OWL have become the favorite representation to encode this knowledge and to store the references to the indexed data. Nowadays, the use of ontologies is omnipresent in all kinds of semantic retrieval tasks in the context of the semantic web [29], as well as in other application contexts such as bioinformatics [113]. Motivated by the lack of meaning in previous IR models, some novel conceptual IR models have been proposed during the last decade, whose main example is the family of ontology-based IR models pioneered by the work of Castells et al. [19], although their origin can be traced back to Rada et al. [119].

An *ontology-based IR model* is any sort of information retrieval model which uses an ontology-based conceptual representation for the content of any sort of information unit, whose main goal is its indexing, retrieval and ranking as regards to a user query. The family of ontology-based IR models can be divided into three groups as follows: (1) the vector ontology-based IR models, such as those introduced by Vallet et al. [140], Fang et al. [35], Castells et al. [19], Mustafa et al. [100], Dragoni et al. [30] and Egozi et al. [31], whose main feature is the use of some adaptation of the standard VSM model to manage concepts instead of meaningless terms; (2) the ontology-based metric space IR models, whose pioneering works are introduced by Rada et al. [119] and subsequently by Lastra-Díaz [58]; and finally, (3) the query-expansion ontology-based IR models, such as those disclosed in patents by Cheslow [21] and Lin et al. [77].

The main features of the family of vector ontology-based IR models, also called adapted-VSM models, are as follows: (1) the use of a conceptual representation for documents and queries based on an ontology; (2) the retrieval of relevant documents through any ontology query language; (3) any sort of vector space for the representation of references to concepts and instances, based on a set of orthogonal base vectors defined by the classes and individuals of the ontology; (4) any sort of adaptation of standard term-frequency weights for the definition of coordinates; (5) the use of cosine function as a ranking method to sort the relevant documents, and finally, (6) a multivector representation and ranking combining different types of features, such as concepts, keywords or ontological features.

2.5.1 Main motivation and hypothesis

Castells et al. [19] and other works in the current family of ontology-based IR models have shown the potential benefits derived from the use of conceptual models as regards their counterpart models based on meaningless terms. However, a carefully study of the underlying assumptions behind most of these conceptual models reveals that there is significant room for improvement in terms of ranking quality, as well as in the precision and recall measures if these underlying aspects were tackled. The main motivation behind most adaptations of classic vector IR models for concept-based semantic search systems has been the definition of a semantic weighting method with the aim of comparing semantically annotated documents.

However, these aforementioned adapted IR models have been using the vector space model (VSM) as a black-box without considering several underlying assumptions of this latter IR model and its consequences.

The main motivation for the research introduced herein is to bridge the gap defined by the main drawbacks of the family of vector ontology-based IR models as detailed below:

1. *Orthogonality condition.* The base vectors of any VSM model are mutually orthogonal, which means that the similarity cosine function between the different base vectors is zero. Thus, two vectors associated with two documents can get a zero, or very low similarity value, when they do not share references to the same concept instances, despite these instances being able to share a common ancestor concept in the taxonomy.
2. *Cardinality mismatch.* Most ontology-based IR models do not include references to classes as sets of objects, and others are mixing references to classes and instances (individuals) at the same representation level. The main idea behind most adaptations of the VSM models to manage the ontology information is to make a mapping from individuals and/or classes to base vectors of the representation vector space. In this way, these IR models are assigning two different and opposite meanings to the same base vector, in one case the base vector represents the occurrence of one object (individuals), whilst in the opposite case, a base vector is representing a collection of objects (classes). These inconsistencies can be summarized as a cardinality mismatch in the adapted VSM models, and the nature of the objects represented by the model.
3. *Statistical fingerprint versus semantic distances.* The metric used to compare documents by most ontology-based IR models is based on the Euclidean angle between normalized vectors (cosine score). The vectors encode the statistical fingerprint of the indexed documents (i.e. the statistical co-occurrence relationships between different concepts in a document), but this metric lacks a meaning in the sense that they are not encoding any semantic distance between concepts, as is done by very well established ontology-based semantic similarity and distance measures. The only exception to this problem is the IR model proposed by Rada et al. [119] which defines a Boolean semantic model, in which the documents are represented by sets of concepts, but the concepts are annotated in binary form without using any semantic weighting method.
4. *Populated ontologies are not directly indexed.* Many vector ontology-based IR models need to retrieve the related documents with the instances and concepts in the query before ranking them. Thus, the populated ontology is not indexed directly else it needs to be searched using any ontology-based query language, such as SPARQL or any other.
5. *Lack of a semantic weighting.* The weights in adapted VSM models are statistical values, not related to the real semantic weight of the concept/instance in the document.

6. *Continuity problems* of some proposed metrics on sets, such as the metric introduced by Rada et al. [119] with the aim of computing the distance between documents. In this latter article, the authors report some continuity problems around close documents. We argue that the source of this discontinuity problem is that the distance function between documents proposed by Rada et al. [119] does not satisfy the coincidence axiom of a metric, thus it is not a well-defined metric on sets.

The aforementioned drawbacks detail a set of intrinsic semantic and geometric structures which are inconsistently encoded in the semantic spaces defined by the current family of vector ontology-based IR models. Thus, our main hypothesis for the research introduced herein is as follows:

Hypothesis 7 (Structure-preserving IR model) *A novel ontology-based IR model which preserves all semantic and geometric structures intrinsically encoded by the base ontology used to index the information could improve the performance of the state-of-the-art IR models in terms of document ranking, precision and recall.*

2.5.2 Research problem and objectives

In order to bridge the gap detailed in the previous section and by evaluating our main hypothesis 7, this research tackles the problem of designing and evaluating a new structure-preserving ontology-based IR model, called *Intrinsic Ontological Spaces*, for the indexing and retrieval of semantically annotated data, such as text documents, web pages, or any sort of information that can be represented as a set of semantic annotations (individuals and classes) in any sort of base ontology. The proposed IR model bridges the gap of modelling inconsistencies in current methods through the integration of the intrinsic structure of any populated ontology in the definition of the representation space itself. The main objectives of the research detailed herein are as follows:

1. To propose a new ontology-based IR model, called *Intrinsic Ontological Spaces*, which is based on the preservation of all semantic and geometric structures intrinsically encoded by the base ontology as intrinsic properties of the resulting semantic space.
2. To evaluate the new ontology-based IR model in an information retrieval task.
3. To evaluate and compare the performance of the new ontology-based IR model with the state-of-the-art methods in the family of vector ontology-based IR models.

2.5.3 Evaluation problems leading to abandoning this task

The building of benchmarks and datasets for the evaluation of semantic search systems has been identified as an urgent need and a line of research in itself. For instance, Fernández et al. [37] introduce a benchmark for the evaluation of semantic search systems based on a TREC dataset, which has been subsequently abandoned.

Another known problem is that most large IR datasets only provide the ranking scores for the set of relevant documents associated to the user queries, instead of providing the scores for the entire corpus. Likewise, Uren et al. [139] surveys the evaluation efforts of a semantic search system and propose “the development of extensible evaluation benchmarks and the use of logging parameters for evaluating individual components of search systems” as main working directions, whilst Elbedweihy et al. [32, abstract] point out that “the evaluation of Semantic Web search systems has largely been developed in isolation from mainstream IR evaluation with a far less unified approach to the design of evaluation activities. This has led to slow progress and low interest when compared to other established evaluation series, such as TREC for IR or OAEI for Ontology Matching”. In this latter paper, the authors also identify the weaknesses of the current semantic search evaluation and “highlight the future need for a more comprehensive evaluation framework that addresses current limitations”.

A first review of the literature led us to be aware of the difficulties for the evaluation of our new ontology-based IR model. These aforementioned difficulties were confirmed by one expert in the area, Dr. Miriam Fernández, research fellow at The Open University. Miriam Fernández¹ introduces the pioneering work in the modern family of vector ontology-based IR models in [19], which is the main contribution of her PhD thesis [38]. She has been involved in development of semantic search systems based on ontologies since the very beginning of the Semantic Web, her being one of its most active researchers. In a series of personal communications and one working meeting with her, she told us that the experiments reported by Castells et al. [19] cannot currently be reproduced because many of the software components required are not currently available. On the other hand, she gave us her wise and timely advice on the current evaluation difficulties of the family ontology-based Information Retrieval (IR) models derived from the lack of well-defined benchmark and datasets for the evaluation of semantic search systems. Likewise, she warned us on the reproducibility problems in this line of research. Thus, we decided to suspend our research activity in this line of research with the aim of focusing on our first line of research into ontology-based semantic similarity measures. However, two recent works on text document similarity, introduced by Benedetti et al. [14] and Ni et al. [101], use the LP50 dataset in the evaluation of their methods. Thus, the experimental setup used by these latter works provide us the possibility of evaluating our new ontology-based IR model proposed in a document similarity task, instead of a semantic search task. For this reason, we plan to resume the research detailed in this section in the mid term.

¹<http://kmi.open.ac.uk/people/member/miriam-fernandez>

Chapter 3

Theoretical Foundations and Methodology

This chapter briefly introduces the theoretical foundations of this thesis, which cover several well-established mathematical theories, such as the theories of graphs, lattices, partially ordered sets (posets) and probability spaces, as well as other theories in cognitive psychology such as the theory of categorization compiled by Margolis and Laurence [84], prototype theory [124], exemplar theory [103] and the theory of geometric conceptual spaces introduced by Gärdenfors [41].

The rest of the chapter is structured as follows. Section 3.1 details the theoretical foundations and closely related lines of research of this thesis, whilst section 3.2 introduces our research methodology.

3.1 Theoretical foundations

Our first line of research (WP1) belongs to the family of ontology-based semantic similarity measures, and more specifically to those based on Information Content (IC) theory. This aforementioned family of semantic measures has been the object of study in many different fields. For instance, in the field of information retrieval, Rada et al. [119] introduce a similarity measure between concepts defined as the length of the shortest path between concepts in a taxonomy for a document retrieval system. In the field of knowledge engineering, Cross and Hu [26] review the use of semantic similarity measures on the ontology alignment (OA), whilst Tversky [138] proposed a feature-based semantic similarity measure in the field of cognitive psychology, and Mazandu et al. [88] review most semantic similarity measures based on Gene Ontology [6] which have been proposed and evaluated in many different tasks in genomics.

The main aim of the family of ontology-based semantic similarity measures is to estimate the human similarity judgments between word and concepts, whose nature and representation are two of the oldest research problems in philosophy and cognitive psychology [83]. Thus, from a cognitive point of view, our research is related to broadly accepted theories in cognitive psychology, such as the theory of categorization, prototype theory and exemplar theory. The theory of categorization is introduced by Rosch and Mervis [123], Rosch et al. [124] and Rosch [122]. In her

pioneering series of works, Eleanor Rosch introduces her classic theory of prototypes in contraposition to the Classic Theory of Concepts in philosophy which defines a concept as a closed set of features. For a detailed analysis of the work of Rosch, we refer the reader to the survey by Peraita Adrados and Labra [112], as well as the critical review of Lakoff [57].

The Classic Theory of concepts comes from a very old tradition in philosophical logic which can be traced back to Socrates and Plato. It argues that most concepts can be defined as closed sets of features which can be verified in a deterministic way. This latter notion means that given any concept defined as a feature set and an unknown individual, it is always possible to find a well-defined criteria, or logic predicate, which could be used to confirm unambiguously the verification of each feature by the individual under study, and thus its belonging to the examined category. Margolis [83] and Laurence and Margolis [69] introduce a critical analysis of the Classic Theory of Concepts and review the evolution of the Theory of Categorization from this classic view to the most modern approaches proposed by the Theory of Prototypes of Rosch and Mervis [123] and its direct evolution known as the Theory of Exemplars introduced by Nosofsky [102, 103] and revised by Smith and Medin [135]. The theory of exemplars argues that the human beings use a particular instance of a category, the so called ‘exemplar’, as base for the comparison and ranking of the degree of belonging of any unknown instance to the examined category. On the other hand, in the theory of prototypes every category is represented by an abstract average instance which is used as base exemplar with the aim of setting the degree of belonging of any unknown instance to the examined category. There is a large corpus of literature on the theory of categorization in the field of cognitive psychology; however, for an introductory reading or in-depth review of the topic, we refer the reader to the collection of works edited by Margolis and Laurence [84] and the more recent book by Margolis and Laurence [85]. Finally, another line of research in cognitive sciences that is closely related to our research, in fact it was a source of inspiration for a couple of IC models introduced by Lastra-Díaz and García-Serrano [59], is the theory of geometric conceptual spaces introduced by Gärdenfors [41].

Most recent research into cognitive sciences has followed a parallel line to the work in the fields of IR and NLP, but it has been more focused on the definition of theoretical models capable of explaining several non-metric phenomena in the human similarity judgments described by Tversky [138] and Pothos et al. [116], such as: (1) asymmetry or non-commutativity, (2) context dependency and (3) the conjunction fallacy. The most recent cognitive similarity model is introduced by Pothos et al. [117] and Pothos and Trueblood [118], being inspired by a quantum probability approach for cognition proposed by Busemeyer and Bruza [18], whose non-commutative nature allows the representation of different non-metric phenomena. However, the quantum probability similarity model has not yet been experimentally evaluated. Other significant contributions to the categorization and prototype theory in the same family of cognitive quantum models are those proposed by Aerts et al. [1].

From a mathematical modeling point of view, the similarity measures are functions defined on taxonomies, or ontologies, which derive from similarity or distance functions defined on these formal structures. Thus, our research is based on many well-founded algebraic theories as follows: the theory of probability spaces [5]; the

theory of metric spaces as detailed in the encyclopedia of the field by Deza and Deza [28]; the closely related theories of partially ordered sets and lattices as proposed by Birkhoff [15], and presented by Lidl and Pilz [76, cp. 1]; valuation metrics on posets such as those reviewed by Monjardet [98] and for semilattices by Ramana Murty and Engelbert [120]; classic set theory; functional analysis, and vector spaces. In addition, our research is inspired by the modern geometric structuralist approach based on the study of the intrinsic properties and invariants of all elements in any mathematical model, which can be traced back to the famous and influential Erlangen program introduced by Klein [55].

Finally, our second line of research belongs to the family of ontology-based information retrieval (IR) models, thus, our research is framed in the field of information retrieval. This line of research shares the same mathematical theories that support our first line of research, and moreover, it is also based on the theory of indexing and information retrieval models, which combines many notions of geometry and statistics, such as vector spaces, machine learning and data mining.

3.2 Research methodology

Our research methodology is defined by the workflow shown in figure 3.1 and detailed in steps 1 to 14 below:

1. Definition of our main research problem.
2. Comprehensive review of the literature on the studied problem as well as other related problems and applications.
3. Synthesis and categorization of the literature based on features such as: strategy and tactics used, functional structure, mathematical models used, application domain, specific problem or motivation, experimental setup, etc.
4. Identification of the gap to be bridged, such as: drawbacks, inconsistencies in the formulation of the models and methods, underlying assumptions, unexplored notions and strategies, formulation of novel hypothesis, refutation of previous conclusions, and study from a novel point of view and disciplines.
5. Proposal of novel methods and hypotheses to bridge the previously identified gap. Correlation and generation of ideas based on analogies and personal intuitions. Inquiry into related ideas in other fields of research, disciplines and related problems.
6. Designing or replication of experiments to evaluate our novel hypotheses and proposals.
7. Implementation of the experiments to evaluate our methods and hypotheses.
8. Replication and reproduction of related methods with the aim of comparing our results with the state of the art.

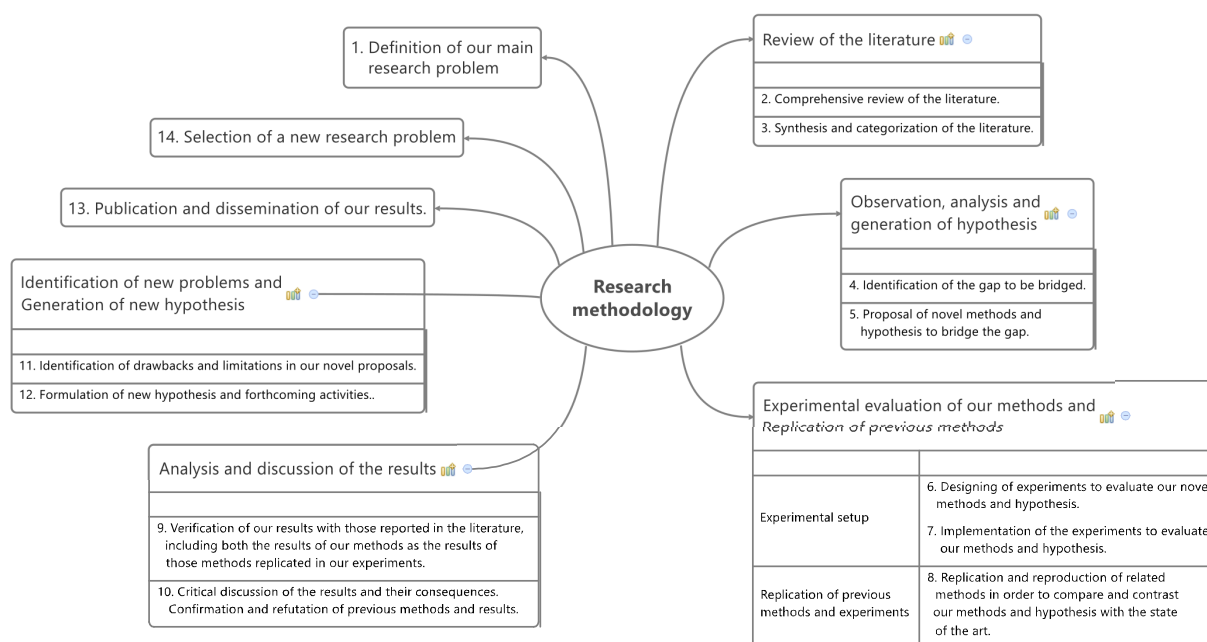


Figure 3.1: Research methodology adopted in this thesis. We place a special emphasis in the replications of previous methods and results, as well as their confirmation and refutation.

9. Verification and contrasting of our results, as well as the results obtained in the replication of other methods and those reported in the literature. Personal communication with the authors whenever it is necessary to clarify any issue for its precise replication and reproduction of their methods and results.
10. Critical discussion of the results and their consequences. Contrast of our results as regards previous methods and results reported in the literature. Confirmation and refutation of previous methods and results reported in the literature based on our own experimentation.
11. Identification of drawbacks and limitations in our novel proposals.
12. Formulation of new hypotheses and forthcoming activities. Identification of potential applications of our methods and results in other related problems or fields of application.
13. Publication and dissemination of our results.
14. Selection of a new research problem from our backlog of new hypotheses, ideas and forthcoming activities and start of a new iteration of our research methodology.

Chapter 4

Conclusions and Future Work

This chapter introduces a summary of the main conclusions derived from the research carried-out in our first line of research into ontology-based semantic similarity measures and IC models based on WordNet (WP1). On the other hand, we abandoned our second line of research (WP2) in our first academic year, as mentioned in section 2.5.3, because of the lack of well-defined benchmarks and datasets for the evaluation of ontology-based IR models. Thus, we will not introduce any conclusions about our research into WP2 herein, beyond confirming the evaluation problems in the field of semantic search.

4.1 Main conclusions

The main conclusions derived from the research introduced by Lastra-Díaz and García-Serrano [60] are as follows:

- 1.1 We introduce one IC-based semantic distance [60, equation 7] and three new IC-based similarity measures [60, equations 8, 11 and 12] based on a generalization and normalization of the classic Jiang-Conrath distance, which outperform the state-of-the-art methods in the RG65 dataset as shown in [60, table 4]. Thus, we positively confirm our hypothesis 1. Our family of weighted Jiang-Conrath similarity measures is subsequently confirmed in a more conclusive manner as the new state of the art of the problem in the largest experimental survey of the field reported in [61, table 12], including their evaluation and comparison on the five most significant datasets and a very detailed statistical significance analysis based on the Spearman correlation metric.
- 1.2 Our $\text{cos}J^{\mathcal{E}C}$ similarity measure, which is a non-linear normalization of the classic Jiang-Conrath similarity measure, in combination with any IC model obtains a higher Pearson correlation value in all word similarity benchmarks than J&C similarity measure (see columns corresponding to both measures in tables 5 to 8 [60]). Likewise, our $\text{cos}wJ^{\mathcal{E}C}$ similarity measure, which is a non-linear normalization of our weighted Jiang-Conrath ($wJ^{\mathcal{E}C}$) similarity measure, in combination with any IC model obtains a higher Pearson correlation value in all word similarity benchmarks than $wJ^{\mathcal{E}C}$ similarity measure (see columns corresponding to both measures in tables 5 to 8 [60]). Thus, we positively confirm our hypothesis 2.

- 1.3 We introduce an up-to-date experimental survey [60, tables 4-8], whose aim is the uniform comparison based on our own software implementation of the most recent and relevant similarity measures based on WordNet, especially the families of IC-based similarity measures and intrinsic IC models.
- 1.4 We introduce an experimental comparison between the intrinsic and corpus-based IC models [60, tables 6-7] that allows some previous conclusions on the outperformance of the intrinsic IC models on the corpus-based to be refuted.
- 1.5 We confirm that the state-of-the-art into similarity measures is lead by the family of IC-based semantic similarity measures [60, table 4], specifically by our new cosine-normalized measures and the Meng and Gu [89] similarity measure.
- 1.6 The use of any hybrid IC-based semantic similarity measure based on the length of the shortest path is refuted because they demand a higher computational cost than other IC-based measures, but they do not show a statistically significant difference as regards these latter ones using the Pearson correlation metric. This conclusion can be drawn by comparing the Pearson correlation values obtained by our *coswJ&C* and Zhou et al. [147] similarity measures with those obtained by our *cosJ&C* similarity measure, as shown in [60, table 4].
- 1.7 Despite the corpus-based IC models evaluated herein obtaining rivaling results as regards the state-of-the-art intrinsic IC models, we confirm that the intrinsic IC models slightly outperform the former ones [60, tables 6-7]. However, the difference between the corpus-based IC models and the intrinsic ones is smaller than that reported in the literature, which was based on corpus-based IC models built with the Resnik method on other WordNet-based frequency files.
- 1.8 We confirm that there is no significant difference in the performance of the ontology-based semantic similarity measures in different versions of WordNet [60, table 4].
- 1.9 We warn of the finding of several reproducibility problems in the replication of several methods and experimental results previously reported in the literature, as well as the discovery of contradictory results [60, section 5.4]. Thus, we invite to the research community to replicate previous methods and experiments in their future research.

The main conclusions derived from the research introduced by Lastra-Díaz and García-Serrano [59] are as follows:

- 2.1 We introduce five new intrinsic IC models and one new corpus-based IC model based on the preservation of the probabilistic structure [59, table 3], and the integration of a notion of cognitive similarity inspired by cognitive evidence.
- 2.2 We show that the proposed approach defines an open framework for the development of new intrinsic IC models based on alternative forms of estimating the conditional probabilities between concepts [59, section 4.2].

- 2.3 Most of our new intrinsic IC models rival the state-of-the-art models, with the exception of the naive *CondProbUniform* model [59, table 4-5].
- 2.4 We show that the integration of the probabilistic structure in the IC models is helpful in getting results rivalling the state-of-the-art IC models, but it is not enough to outperform the state of the art by itself [59, table 4]. Thus, our hypothesis 3 is not positively confirmed. However, we expect that the encoding of the structure axioms into the IC models contributes to a better understanding of the problem, as well as the start of a line of research into conditional probability estimation.
- 2.5 On the other hand, the results of the *CondProbCosine* and *CondProbLogistic* model confirm that the encoding of cognitive similarity notions within the IC models and measures is a line of research that is worth exploring [59, fig. 2]. However, our hypothesis 4 is not positively confirmed.
- 2.6 We show that there is no a statistically significant difference between most intrinsic IC models and IC-based similarity measures and the baselines of the experiments [59, figs. 2-3].
- 2.7 Despite the IC model introduced by Seco et al. [134] obtaining the highest overall average correlation values [59, table 5], the statistical evidence shows that the IC model introduced by Sánchez and Batet [128] obtains a significant statistical outperformance over the baseline and the rest of the IC models [59, fig. 2], this latter model being the IC model that best generalizes any IC-based similarity measure. The Sánchez and Batet [128] IC model is the only capable of statistically outperforming the corpus-based IC model defined as baseline [59, fig. 2].
- 2.8 We show that our *cosJ&C* semantic similarity measure introduced in [60] obtains the best overall results [59, table 6], obtaining a statistically significant outperformance over the rest of the IC models and measures in comparison with the baseline [59, fig. 3]. However, a more in-depth confidence interval analysis between the similarity measures introduced by Pirró and Euzenat [115] (FaITH), Meng and Gu [89] and Lastra-Díaz and García-Serrano [60] (cosJ&C) confirms that there is no a statistically significant difference between them.
- 2.9 The lack of a statistically significant difference between most intrinsic IC models and the corpus-based IC model *Resnik_{ic-treebank-add1}* defined as the baseline [59, fig. 2] allows the following conclusions to be extracted: (1) this fact refutes a previous belief about the outperformance of the intrinsic IC models over the corpus-based ones, confirming the same finding in our aforementioned work [60], and (2) this fact confirms the achievements of the family of intrinsic IC models, which offers a practical alternative to the corpus-based models without a significant reduction in performance.
- 2.10 Among the set of rivaling state-of-the-art intrinsic IC models we have those introduced by Seco et al. [134], Yuan et al. [144], Meng et al. [90], Sánchez

and Batet [128], Sánchez et al. [129], Harispe et al. [48], *CondProbCosine*, *CondProbHypo*, and *CondProbLeaves* [59, fig. 2].

- 2.11 The statistical significance of the results confirms that most of the IC models offer similar results, and the problem is still open [59, fig. 2].

The main conclusions derived from the research introduced by Lastra-Díaz and García-Serrano [61] are as follows:

- 3.1 We introduce a refinement of our recent family of well-founded Information Content models introduced in [59], eight new intrinsic IC models and one new corpus-based IC model [61, section 4] together with a very detailed experimental survey into WordNet based on our software implementation of most methods reported in the literature [61, section 5-6].
- 3.2 We show that the proposed refinement improves the performance of our family of well-founded IC models, and six of our new IC models obtain rivaling results as regard the state-of-the-art intrinsic IC models [61, table 10], making the new *CondProbRefHyponyms* and *CondProbRefCosine* IC models our best performing IC models.
- 3.3 We show that most refined IC models proposed in [61, section 4] outperform in a statistically significant manner their non-refined counterpart IC models. Thus, our hypothesis 5 is positively confirmed. Looking at table 13 in [61], we see that the new IC models *CondProbRefUniform*, *CondProbRefLeaves*, *CondProbRefCosine* and *CondProbRefCorpus*, obtain a statistically significant higher average Spearman correlation than their corresponding non-refined IC models *CondProbUniform*, *CondProbLeaves*, *CondProbCosine* and *CondProbCorpus*. However, the *CondProbRefHyponyms* and *CondProbRefLogistic* IC models are not able to obtain a statistically significant higher performance than their corresponding models *CondProbHyponyms* and *CondProbLogistic*.
- 3.4 The intrinsic IC models introduced by Sánchez et al. [129] and Seco et al. [134] set the state of the art for the family of intrinsic IC models in a statistically significant manner in combination with our *coswJ&C* similarity measure proposed in [60] and those introduced by Zhou et al. [147] (see tables 12 and 14 [61]).
- 3.5 The intrinsic IC models introduced by Seco et al. [134], Sánchez et al. [129] and Yuan et al. [144] are the only ones that statistically outperform the best performing corpus-based IC model used as baseline (see table 10 and first column in table 13 [61]). However, we show that there is no statistically significant difference between most intrinsic IC models and the corpus-based Resnik IC model defined as baseline (see first column in table 13 [61]). Therefore, the aforementioned set of intrinsic IC models can be considered as a practical alternative to the corpus-based ones, and they should be selected in accordance with the IC-based similarity measure used.

- 3.6 The detailed experiment survey carried-out herein allows a very significant conclusion to be drawn: despite the research effort made during the last decade, the IC model introduced by Seco et al. [134] is still, on average, the state of the art (see first row in table 10 [61]).
- 3.7 The new state-of-the-art in intrinsic IC models and intrinsic IC-based similarity measures is set out in a statistically significant manner by the Sánchez et al. [129] IC model in combination with our *coswJ&C* similarity measure [60], and the Seco et al. [134] IC model in combination with the Zhou et al. [147] similarity measure (see first two rows in table 12 and table 14 [61]). The statistical significance of data is based on the Spearman correlation metric, and this latter conclusion allows our hypothesis 1, as well as the conclusions introduced by Lastra-Díaz and García-Serrano [60] as regards our novel family of IC-based similarity measures, to be positively confirmed in a more conclusive manner.
- 3.8 The set of classic IC-based similarity measures, defined by the Resnik, Lin and Jiang-Conrath measures, have also been definitively outperformed in a statistically significant manner by a small set of IC-based similarity measures developed during the last decade, among which we find the similarity measures introduced by Zhou et al. [147] and our *coswJ&C* similarity measure introduced in [60] (see table 12 and the columns corresponding to the Resnik, Lin and Jiang-Conrath measures in table 14 [61]).
- 3.9 In addition, the classic Jiang-Conrath similarity measure and its two monotone transformations, our *cosJ&C* measure and the similarity measure introduced by Garla and Brandt [42], statistically outperform the Resnik and Lin similarity measures, whilst our *cosJ&C* similarity measure obtains a statistically significant higher average Pearson correlation value than the J&C similarity measure (see table 12 in [61] and figure 2 in [59]). However, we show that there is no a statistically significant difference between the two aforementioned pairs of outperforming IC-based similarity measures [61, table 14].
- 3.10 Despite our *coswJ&C* similarity measure and the Zhou et al. [147] measure setting the state of the art of the problem, their computational cost prevent their practical use in comparison with other measures [61, table 9], such as our *cosJ&C* similarity measure [60] and the Hadj Taieb et al. [45] measure. However, there is no a statistically significant difference between the two latter aforementioned measures [61, table 14]. Thus, the *cosJ&C* and Hadj Taieb et al. [45] measures are, statistically speaking, the best option from the aforementioned set of similarity measures with a practical computational cost.
- 3.11 We show that the state of the art in the family of ontology-based similarity measures and concept similarity models is led by the family of IC-based measures, more specifically by our *coswJ&C* similarity measure and the Zhou et al. [147] measure, both derived from the Jiang-Conrath similarity measure.
- 3.12 Finally, we have made another significant finding. Contrary to the common belief among the research community, only a small set of state-of-the-art hy-

brid IC-based similarity measures derived from the J&C measure obtain a statistically significant higher average Spearman correlation value than the family of path-based similarity measures (see columns corresponding to the path-based measures in table 14 [61]), a fact that explains some unexpected results in applications based on similarity measures reported in the literature, such as that reported by Alonso and Contreras [4, section 8].

The main conclusions derived from the research introduced by Lastra-Díaz et al. [68] are as follows:

- 4.1 We introduce a new and linearly scalable representation model for large taxonomies, called *PosetHERep*, and the *HESML V1R2* [63] semantic measures library based on the former [68, section 3].
- 4.2 Most HESML V1R2 algorithms exhibit linear complexity (see tables 20 and 21, and figure 3 [68]), thus they are linearly scalable as predicted by our theoretical analysis.
- 4.3 We show in a statistically significant manner that HESML V1R2 is the most efficient and scalable publicly available software library of ontology-based similarity measures and intrinsic IC models based on WordNet, outperforming SML [48] and WNetSS [13] by several orders of magnitude in most benchmarks (see tables 19 to 21 and figure 3 [68]).
- 4.4 There is no a statistically significant difference in the performance of HESML and SML in the evaluation of a classic IC-based similarity measure based on WordNet (see p-value in table 19 [68]), unlike the evaluation of any path-based semantic similarity measure in which HESML is much more efficient (see last two columns in table 20 and figure 3.i [68]).
- 4.5 The performance of SML in the evaluation of path-based semantic similarity prevents its usage in the evaluation of this type of measures on large taxonomies as WordNet (see average running times for medium-size taxonomies in last column of table 20 [68]).
- 4.6 We show that *PosetHERep* and *HESML*, conversely to common belief, are able to improve the performance and scalability of the state-of-the-art semantic measures libraries significantly without caching using a proper intrinsic representation model for taxonomies. HESML significantly outperforms SML in those methods in which SML uses caching, such as the retrieval of the LCA [68, fig. 3.f] and MICA vertexes [68, fig. 3.g], and the set of subsumed leaves of a vertex [68, fig. 3.h].
- 4.7 The performance of WNetSS is more than three orders of magnitude lower than HESML and SML because of its caching strategy based on a relational database [68, table 19].
- 4.8 The overall outperformance of HESML on SML proves our main hypothesis 6 and answers our two main research questions 1 and 2 positively. Thus, our

results allow the following conclusions to be drawn: (1) a new intrinsic representation model for taxonomies like that proposed by PosetHERep is able to improve the performance and scalability of the state-of-the-art semantic measures libraries significantly; and (2) it is possible to improve the performance and scalability of the state-of-the-art semantic measures libraries significantly without using any caching strategy by using the PosetHERep model.

- 4.9 We introduce a set of reproducible experiments based on ReproZip [68, section 4], publicly available at [67], and *HESML*, which allow our experimental surveys introduced in [60, 59, 61] to be reproduced exactly.
- 4.10 We introduce, for the first time, a replication framework and dataset called *WNSimRep v1* [68, section 5] which is publicly available at [66].
- 4.11 We introduce, for the first time, a benchmark of semantic measures libraries, publicly available at [64], which allows the benchmark introduced by [68, section 6] to be reproduced exactly.

Finally, table 4.1 shows a summary on the confirmation of the main hypotheses and research questions studied by this thesis.

4.2 Future work

As forthcoming activities, we plan to continue our work in three complementary directions as follows:

1. *Functional extension of our HESML software library.* We plan to extend *HESML* in order to support Wikidata [141] and non “is-a” relationships in the short term, whilst in the mid term, we expect to support the Gene Ontology (GO), MeSH and SNOMED-CT ontologies. In addition, we plan to include further ontology-based similarity measures and IC models reported in the literature, as well as the possibility of importing word embedding files with the aim of allowing the experimental comparison of state-of-the-art ontology-based and corpus-based similarity measures and methods. Finally, we plan to extend *HESML* in the mid- and long-term with the aim of exploring its scalability and performance in graph mining of very large graphs (billions of nodes) based on a single computer. Our current intuition is that an extended version of *HESML* for general graphs could be able to compete with state-of-the-art centralized graph mining libraries such as GraphChi [56], TurboGraph [47] and MMAP [79], as well as other distributed graph mining libraries such as PREGEL [81] which are evaluated by Batarfi et al. [8].
2. *Evaluation of our ontology-based IR models introduced in [58] and disclosed in [65].* As we mentioned in section 2.5, we decided to suspend this line of research because of the well-known lack of well-defined datasets and benchmarks for the evaluation of ontology-based information retrieval models, as well as the huge difficulties for their construction. However, two recent works on text document similarity introduced by Benedetti et al. [14] and Ni et al. [101] have used

Id	Hypothesis	Results
H1	A new semantic distance defined as the length of the shortest path between concepts in a weighted taxonomy whose edge weights are set to the IC value of the conditional probability between its child and parent concepts, or the difference in absolute value of their IC values, could improve the estimation of the human similarity judgements between words and concepts obtained by the current state-of-the-art similarity measures.	Positively confirmed
H2	A new semantic similarity measure defined as a proper non-linear normalization of the classic Jiang-Conrath distance, or our new weighted Jiang-Conrath distance, could improve the estimation of the human similarity judgements between words and concepts obtained by their non-normalized versions.	Positively confirmed
H3	A new family of intrinsic and corpus-based IC models based on the explicit encoding of the structure axioms derived from the conditional probabilities could improve the performance obtained by current state-of-the-art IC models in semantic similarity tasks, especially to provide a better understanding of the problem.	Rivaling results, not confirmed
H4	A new family of intrinsic and corpus-based IC models based on the explicit encoding of a non-linear cognitive distance between parent and child concepts could improve the performance obtained by current state-of-the-art IC models in semantic similarity tasks.	Rivaling results, not confirmed
H5	The solution to the two aforementioned structural drawbacks of the main computation algorithm of our family of well-founded IC models could lead us to an improvement in their performance, in addition to fixing an algebraic inconsistency that moves the family of well-founded IC models away from their original design principles.	Positively confirmed
H6	A new representation model for taxonomies which properly encodes their intrinsic structure, together with a new software library based on it, should bridge the aforementioned gap of scalability and performance of the current state-of-the-art semantic measures libraries.	Positively confirmed
Q1	Is a new intrinsic representation model for taxonomies able to improve the scalability and performance of the current state-of-the-art semantic measures libraries significantly?	Positively answered
Q2	Is it possible to improve the scalability and performance of the state-of-the-art semantic measures libraries without using any caching strategy significantly?	Positively answered
H7	A novel ontology-based IR model which preserves all semantic and geometric structures intrinsically encoded by the base ontology used to index the information could improve the performance of the state-of-the-art IR models in terms of document ranking, precision and recall.	Not answered

Table 4.1: Results obtained for the main hypotheses and research questions studied by this thesis.

the LP50 dataset in the evaluation of their methods. The LP50 dataset is introduced by Lee et al. [72], and it is made up of 50 text documents together with all the human similarity judgements between each pair of documents which allows standard Pearson and Spearman correlation metrics to be used to evaluate the quality of the document similarity measures proposed. Thus, we think that this well-defined document similarity task could be used in order to evaluate our ontology-based IR models, instead of other classic IR tasks based in queries and sets of relevant documents, such as those proposed at TREC conferences.

3. *Exploration of different applications in genomics.* Mazandu et al. [88] introduce a recent survey on the family of ontology-based semantic similarity measures based on Gene Ontology (GO) [6] and its multiple applications in genomics. Thus, we plan to explore this line of research in the long term, once HESML supports the GO ontology.

Chapter 5

Scientific contributions

This chapter sets out all contributions derived directly from this thesis, which are divided into five types as follows: (1) peer-reviewed articles, (2) technical reports, (3) patent applications, (4) software libraries and (5) replication datasets and benchmarks.

5.1 Peer-reviewed articles

1. Lastra-Díaz, J. J., & García-Serrano, A. (2015). A novel family of IC-based similarity measures with a detailed experimental survey on WordNet. *Engineering Applications of Artificial Intelligence Journal*, 46, 140–153.
<http://dx.doi.org/10.1016/j.engappai.2015.09.006>
2. Lastra-Díaz, J. J., & García-Serrano, A. (2015). A new family of information content models with an experimental survey on WordNet. *Knowledge-Based Systems*, 89, 509–526.
<http://dx.doi.org/10.1016/j.knosys.2015.08.019>
3. Lastra-Díaz, J. J., García-Serrano, A., Batet, M., Fernández, M., & Chirigati, F. (2017). HESML: a scalable ontology-based semantic similarity measures library with a set of reproducible experiments and a replication dataset. *Information Systems*, 66, 97–118.
<http://dx.doi.org/10.1016/j.is.2017.02.002>

5.2 Technical reports

1. Lastra-Díaz, J. J., & García-Serrano, A. (2016). A refinement of the well-founded Information Content models with a very detailed experimental survey on WordNet (No. TR-2016-01). NLP and IR Research Group. ETSI Informática. Universidad Nacional de Educación a Distancia (UNED).
[61]

5.3 Patent applications

1. Lastra Díaz, J. J., & García Serrano, A. (2016). System and method for the indexing and retrieval of semantically annotated data using an ontology-based information retrieval model. United States Patent and Trademark Office (USPTO) Application, US2016/0179945 A1.

5.4 Software libraries

1. Lastra-Díaz, J. J., & García-Serrano, A. (2016). HESML V1R2 Java software library of ontology-based semantic similarity measures and information content models. Mendeley Data, v2.
<http://dx.doi.org/10.17632/t87s78dg78.2>
2. Lastra-Díaz, J. J., & García-Serrano, A. (2016). HESML V1R1 Java software library of ontology-based semantic similarity measures and information content models. Mendeley Data v1.
<http://dx.doi.org/10.17632/t87s78dg78.1>

5.5 Replication datasets and benchmarks

1. Lastra-Díaz, J. J., & García-Serrano, A. (2016). WNSimRep: a framework and replication dataset for ontology-based semantic similarity measures and information content models. Mendeley Data v1.
<http://dx.doi.org/10.17632/mpr2m8pycs.1>
2. Lastra-Díaz, J. J., & García-Serrano, A. (2016). WordNet-based word similarity reproducible experiments based on HESML V1R1 and ReproZip. Mendeley Data, v1.
<http://dx.doi.org/10.17632/65pxgskhz9.1>
3. Lastra-Díaz, J. J., & García-Serrano, A. (2016, November). HESML_vs_SML: scalability and performance benchmarks between the HESML V1R2 and SML 0.9 semantic measures libraries. Mendeley Data, v1.
<http://dx.doi.org/10.17632/5hg3z85wf4.1>

Chapter 6

Impact factor of the publications

Table 6.1 shows the JCR quartile and 2-year Impact Factor (IF) of our three main publications corresponding to the JCR-2015 and JCR-2016 rankings, as shown in figures 6.1 and 6.2 respectively.

The JCR-2016 ranking has been recently published by Thomson at Web of Science (WoS) InCites portal; however, it was not available in the Web of Knowledge (WoK) JCR analytics tool in the time of preparation of this manuscript.

Reference	Journal	2-year IF		Quartile
		2015	2016	
Lastra-Díaz and García-Serrano [60]	Engineering Applications of Artificial Intelligence	2.368	2.894	Q1
Lastra-Díaz and García-Serrano [59]	Knowledge-Based Systems	3.325	4.529	Q1
Lastra-Díaz et al. [68]	Information Systems	1.832	2.777	Q1
Overall Impact Factor		7.525	10.2	

Table 6.1: 2-year JCR impact factors of the three main publications derived from this thesis. All of the above journals are edited by Elsevier.



	Title20	Year	IMPACT_FACTOR	CATEGORY_RANKING
1	ENG APPL ARTIF INTEL	2015	2,368	Q1
2	ENG APPL ARTIF INTEL	2015	2,368	Q1
3	ENG APPL ARTIF INTEL	2015	2,368	Q1
4	ENG APPL ARTIF INTEL	2015	2,368	Q1
5	INFORM SYST	2015	1,832	Q1
6	KNOWL-BASED SYST	2015	3,325	Q1

Figure 6.1: JCR-2015 Impact Factor and Quartile of our three main publications (source: WoK-FECYT)

Journal Titles Ranked by Impact Factor

Compare Selected Journals		Add Journals to New or Existing List		
Select All		Full Journal Title	Total Cites	Journal Impact Factor ▼
<input type="checkbox"/>	1	KNOWLEDGE-BASED SYSTEMS	7,763	4.529
<input type="checkbox"/>	2	ENGINEERING APPLICATIONS OF ARTIFICIAL INTELLIGENCE	5,111	2.894
<input type="checkbox"/>	3	INFORMATION SYSTEMS	2,044	2.777

Figure 6.2: JCR-2016 impact factor of our three main publications (source: WoS InCites)

Bibliography

- [1] Aerts, D., Broekaert, J., Gabora, L., Sozzo, S., 30 Mar. 2016. Generalizing Prototype Theory: A Formal Quantum Framework. *Frontiers in psychology* 7, 418.
- [2] Agirre, E., Alfonseca, E., Hall, K., Kravalova, J., Paşca, M., Soroa, A., 2009. A Study on Similarity and Relatedness Using Distributional and WordNet-based Approaches. In: *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics. NAACL '09*. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 19–27.
- [3] Al-Mubaid, H., Nguyen, H. A., Jul. 2009. Measuring Semantic Similarity Between Biomedical Concepts Within Multiple Ontologies. *IEEE transactions on systems, man and cybernetics. Part C, Applications and reviews: a publication of the IEEE Systems, Man, and Cybernetics Society* 39 (4), 389–398.
- [4] Alonso, I., Contreras, D., Feb. 2016. Evaluation of semantic similarity metrics applied to the automatic retrieval of medical documents: an UMLS approach. *Expert Systems with Applications* 44, 386–399.
- [5] Ash, R. B., Doléans-Dade, C. A., 2000. *Probability & Measure Theory*, 2nd Edition. Academic Press.
- [6] Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Michael Cherry, J., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., Sherlock, G., 1 May 2000. Gene Ontology: tool for the unification of biology. *Nature genetics* 25 (1), 25–29.
- [7] Basili, R., Pennacchiotti, M., 2010. Distributional lexical semantics: Toward uniform representation paradigms for advanced acquisition and processing tasks. *Natural Language Engineering* 16, 347–358.
- [8] Batarfi, O., El Shawi, R., Fayoumi, A. G., Nouri, R., Beheshti, S.-M.-R., Barnawi, A., Sakr, S., 24 Jul. 2015. Large scale graph processing systems: survey and an experimental evaluation. *Cluster computing* 18 (3), 1189–1213.
- [9] Batet, M., 2011. A study on semantic similarity and its application to clustering: Enabling the classification of textual data. VDM Verlag.

- [10] Batet, M., 2011. Ontology-based semantic clustering. *AI Communications. The European Journal on Artificial Intelligence* 4 (3), 291–292.
- [11] Batet, M., Erola, A., Sánchez, D., Castellà-Roca, J., 1 Sep. 2013. Utility preserving query log anonymization via semantic microaggregation. *Information sciences* 242, 49–63.
- [12] Batet, M., Sánchez, D., Valls, A., Feb. 2011. An ontology-based measure to compute semantic similarity in biomedicine. *Journal of Biomedical Informatics* 44 (1), 118–125.
- [13] Ben Aouicha, M., Taieb, M. A. H., Ben Hamadou, A., 21 Nov. 2016. SISR: System for integrating semantic relatedness and similarity measures. *Soft Computing*, 1–25 <http://dx.doi.org/10.1007/s00500-016-2438-x>.
- [14] Benedetti, F., Beneventano, D., Bergamaschi, S., 24 Oct. 2016. Context Semantic Analysis: A Knowledge-Based Technique for Computing Inter-document Similarity. In: *Similarity Search and Applications*. Springer, Cham, pp. 164–178.
- [15] Birkhoff, G., 1967. *Lattice Theory*, 3rd Edition. Vol. XXV of Colloquium Publications. American Mathematical Society.
- [16] Budanitsky, A., Hirst, G., 2001. Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. In: *Workshop on WordNet and Other Lexical Resources, Second Meeting of the North American Chapter of the Association for Computational Linguistics*. Vol. 2. ACL, Pittsburgh, PA, pp. 29–34.
- [17] Budanitsky, A., Hirst, G., Mar. 2006. Evaluating WordNet-based Measures of Lexical Semantic Relatedness. *Computational Linguistics* 32 (1), 13–47.
- [18] Busemeyer, J. R., Bruza, P. D., 2012. *Quantum models of cognition and decision*. Cambridge University Press.
- [19] Castells, P., Fernández, M., Vallet, D., 2007. An adaptation of the vector-space model for ontology-based information retrieval. *IEEE transactions on knowledge and data engineering* 19 (2), 261–272.
- [20] Chaves-González, J. M., Martínez-Gil, J., Jan. 2013. Evolutionary algorithm based on different semantic similarity functions for synonym recognition in the biomedical domain. *Knowledge-Based Systems* 37, 62–69.
- [21] Cheslow, R. D., 30 Oct. 2012. System and method for semantic search. United States Patent and Trademark Office (USPTO) application US8301633B2, <https://www.google.com/patents/US8301633>.
- [22] Chirigati, F., Rampin, R., Shasha, D., Freire, J., 2016. ReproZip: computational reproducibility with ease. In: *Proceedings of the 2016 ACM SIGMOD International Conference on Management of Data (SIGMOD)*. Vol. 16. big-data.poly.edu, pp. 2085–2088.

- [23] Clark, S., 2012. Vector space models of lexical meaning. In: Lappin, S., Fox, C. (Eds.), *Handbook of Contemporary Semantics Theory*, 2nd Edition. WILEY Blackwell, Malden, MA, Ch. 16, pp. 493–522.
- [24] Couto, F. M., Pinto, H. S., Oct. 2013. The next generation of similarity measures that fully explore the semantics in biomedical ontologies. *Journal of Bioinformatics and Computational Biology* 11 (5), 1371001.
- [25] Couto, F. M., Silva, M. J., Coutinho, P. M., Apr. 2007. Measuring semantic similarity between Gene Ontology terms. *Data & Knowledge Engineering* 61 (1), 137–152.
- [26] Cross, V., Hu, X., 2011. Using Semantic Similarity in Ontology Alignment. In: *Proc. of the Sixth International Workshop on Ontology Matching (OM), 10th Int. Semantic Web Conference (ISWC 2011)*. Bonn Germany, pp. 61–72.
- [27] Dagher, G. G., Fung, B. C. M., Jul. 2013. Subject-based semantic document clustering for digital forensic investigations. *Data & Knowledge Engineering* 86, 224–241.
- [28] Deza, M., Deza, E., 2009. *Encyclopedia of distances*. Springer.
- [29] Ding, L., Kolari, P., Ding, Z., Avancha, S., Jan. 2007. Using Ontologies in the Semantic Web: A Survey. In: Sharman, R., Kishore, R., Ramesh, R. (Eds.), *Ontologies*. Vol. 14 of *Integrated Series in Information Systems*. Springer US, pp. 79–113.
- [30] Dragoni, M., Da Costa Pereira, C., Tettamanzi, A. G. B., 1 Jan. 2010. An Ontological Representation of Documents and Queries for Information Retrieval Systems. In: *Trends in Applied Intelligent Systems. Lecture Notes in Computer Science*. Springer Berlin Heidelberg, pp. 555–564.
- [31] Egozi, O., Markovitch, S., Gabrilovich, E., 2011. Concept-based information retrieval using explicit semantic analysis. *ACM Transactions on Information Systems (TOIS)* 29, 8.
- [32] Elbedweihi, K. M., Wrigley, S. N., Clough, P., Ciravegna, F., 2014. An Overview of Semantic Search Evaluation Initiatives. *Web Semantics: Science, Services and Agents on the World Wide Web*.
- [33] Erk, K., 2012. Vector Space Models of Word Meaning and Phrase Meaning: A Survey. *Language and linguistics compass* 6, 635–653.
- [34] Fährndrich, J., Weber, S., Ahrndt, S., 27 Sep. 2016. Design and Use of a Semantic Similarity Measure for Interoperability Among Agents. In: Klusch, M., Unland, R., Shehory, O., Pokahr, A., Ahrndt, S. (Eds.), *Multiagent System Technologies. Lecture Notes in Computer Science*. Springer International Publishing, pp. 41–57.
- [35] Fang, W.-D., Zhang, L., Wang, Y.-X., Dong, S.-B., 2005. Toward a semantic search engine based on ontologies. In: *Proceedings of International Conference on Machine Learning and Cybernetics*. Vol. 3. IEEE, pp. 1913–1918.

- [36] Fares, M., Kutuzov, A., Oepen, S., Veldal, E., 23 May 2017. Word vectors, reuse, and replicability: Towards a community repository of large-text resources. In: *Proceedings of the 21st Nordic Conference of Computational Linguistics*. Linköping University Electronic Press, Cothenburg, Sweden, pp. 271+276.
- [37] Fernández, M., López, V., Sabou, M., Uren, V., Vallet, D., Motta, E., Castells, P., 20 Apr. 2009. Using TREC for cross-comparison between classic IR and ontology-based search models at a Web scale. In: *Semantic Search 2009 Workshop at the 18th International World Wide Web Conference (WWW 2009)*,. Madrid, Spain.
- [38] Fernández Sánchez, M., 2009. *Semantically enhanced Information Retrieval: an ontology-based approach*. Ph.D. thesis, Universidad Autónoma de Madrid, Departamento de Ingeniería Informática.
- [39] Fernando, S., Stevenson, M., 2008. A semantic similarity approach to paraphrase detection. In: *Proc. of the 11th Annual Research Colloquium of the UK Special-interest group for Computational Linguistics*. Oxford, UK, pp. 45–52.
- [40] Gao, J. B., Zhang, B. W., Chen, X. H., Mar. 2015. A WordNet-based semantic similarity measurement combining edge-counting and information content theory. *Engineering Applications of Artificial Intelligence* 39, 80–88.
- [41] Gärdenfors, P., 2014. *The Geometry of Meaning: Semantics Based on Conceptual Spaces*. MIT Press.
- [42] Garla, V. N., Brandt, C., 10 Oct. 2012. Semantic similarity in the biomedical domain: an evaluation across knowledge sources. *BMC bioinformatics* 13:261.
- [43] Grego, T., Couto, F. M., 2 May 2013. Enhancement of chemical entity identification in text using semantic similarity validation. *PloS one* 8 (5), e62984.
- [44] Guzzi, P. H., Mina, M., Guerra, C., Cannataro, M., Sep. 2012. Semantic similarity analysis of protein data: assessment with biological features and issues. *Briefings in bioinformatics* 13 (5), 569–585.
- [45] Hadj Taieb, M. A., Ben Aouicha, M., Ben Hamadou, A., Nov. 2014. Ontology-based approach for measuring semantic similarity. *Engineering Applications of Artificial Intelligence* 36, 238–261.
- [46] Hadj Taieb, M. A., Ben Aouicha, M., Bourouis, Y., 22 Jun. 2015. FM3S: Features-Based Measure of Sentences Semantic Similarity. In: Onieva, E., Santos, I., Osaba, E., Quintián, H., Corchado, E. (Eds.), *Proceedings of the 10th International Conference on Hybrid Artificial Intelligent Systems (HAIS 2015)*. Vol. 9121 of LNCS. Springer, Bilbao, Spain, pp. 515–529.
- [47] Han, W.-S., Lee, S., Park, K., Lee, J.-H., Kim, M.-S., Kim, J., Yu, H., 2013. TurboGraph: A Fast Parallel Graph Engine Handling Billion-scale Graphs in a Single PC. In: *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '13*. ACM, New York, NY, USA, pp. 77–85.

- [48] Harispe, S., Ranwez, S., Janaqi, S., Montmain, J., 18 Jun. 2014. The Semantic Measures Library: Assessing Semantic Similarity from Knowledge Representation Analysis. In: Métais, E., Roche, M., Teisseire, M. (Eds.), Proc. of the 19th International Conference on Applications of Natural Language to Information Systems (NLDB 2014). Vol. 8455 of LNCS. Springer, Montpellier, France, pp. 254–257.
- [49] Harispe, S., Ranwez, S., Janaqi, S., Montmain, J., May 2015. Semantic Similarity from Natural Language and Ontology Analysis. Vol. 8 of Synthesis Lectures on Human Language Technologies. Morgan & Claypool publishing.
- [50] Harris, Z. S., 1981. Distributional Structure. In: Hiž, H. (Ed.), Papers on Syntax. Vol. 14 of Synthese Language Library. Springer Netherlands, pp. 3–22.
- [51] Hill, F., Reichart, R., Korhonen, A., 2015. SimLex-999: Evaluating Semantic Models with (Genuine) Similarity Estimation. *Computational Linguistics* 41 (4), 665–695.
- [52] Hirst, G., St-Onge, D., 1998. Lexical chains as representations of context for the detection and correction of malapropisms. In: Fellbaum, C. (Ed.), WordNet: An electronic lexical database. Massachusetts Institute of Technology, pp. 305–332.
- [53] Jiang, J. J., Conrath, D. W., 1997. Semantic similarity based on corpus statistics and lexical taxonomy. In: Proceedings of International Conference Research on Computational Linguistics (ROCLING X). pp. 19–33.
- [54] Jiang, Y., Bai, W., Zhang, X., Hu, J., 2017. Wikipedia-based information content and semantic similarity computation. *Information Processing & Management* 53 (1), 248–265.
- [55] Klein, F., Jul. 1893. A comparative review of recent researches in geometry (translation the german paper published in Erlangen, 1872). *Bulletin of the American Mathematical Society*, 215–249.
- [56] Kyrola, A., Blelloch, G., Guestrin, C., 2012. GraphChi: large-scale graph computation on just a PC. In: Presented as part of the 10th USENIX Symposium on Operating Systems Design and Implementation (OSDI 12). usenix.org, pp. 31–46.
- [57] Lakoff, G., 1999. Cognitive models and prototype theory. In: Margolis, E., Laurence, S. (Eds.), *Concepts: Core Readings*. MIT Press, pp. 391–421.
- [58] Lastra-Díaz, J. J., 29 Sep. 2014. Intrinsic Semantic Spaces for the representation of documents and semantic annotated data. Master’s thesis, Universidad Nacional de Educación a Distancia (UNED). Department of Computer Languages and Systems, <http://e-spacio.uned.es/fez/view/bibliuned:master-ETSIInformatica-LSI-Jlastra>.

- [59] Lastra-Díaz, J. J., García-Serrano, A., Nov. 2015. A new family of information content models with an experimental survey on WordNet. *Knowledge-Based Systems* 89, 509–526.
- [60] Lastra-Díaz, J. J., García-Serrano, A., Nov. 2015. A novel family of IC-based similarity measures with a detailed experimental survey on WordNet. *Engineering Applications of Artificial Intelligence Journal* 46, 140–153.
- [61] Lastra-Díaz, J. J., García-Serrano, A., 7 Jul. 2016. A refinement of the well-founded Information Content models with a very detailed experimental survey on WordNet. Tech. Rep. TR-2016-01, NLP and IR Research Group. ETSI Informática. Universidad Nacional de Educación a Distancia (UNED), <http://e-spacio.uned.es/fez/view/bibliuned:DptoLSI-ETSI-Infornes-Jlastra-refinement>.
- [62] Lastra-Díaz, J. J., García-Serrano, A., 2016. HESML V1R1 Java software library of ontology-based semantic similarity measures and information content models. Mendeley Data, v1, <http://dx.doi.org/10.17632/t87s78dg78.1>.
- [63] Lastra-Díaz, J. J., García-Serrano, A., 2016. HESML V1R2 Java software library of ontology-based semantic similarity measures and information content models. Mendeley Data, v2, <http://dx.doi.org/10.17632/t87s78dg78.2>.
- [64] Lastra-Díaz, J. J., García-Serrano, A., Nov. 2016. HESML_vs_SML: scalability and performance benchmarks between the HESML V1R2 and SML 0.9 semantic measures libraries. Mendeley Data, v1, <http://dx.doi.org/10.17632/5hg3z85wf4.1>.
- [65] Lastra Díaz, J. J., García Serrano, A., 23 Jun. 2016. System and method for the indexing and retrieval of semantically annotated data using an ontology-based information retrieval model. United States Patent and Trademark Office (USPTO) application US2016/0179945 A1, <https://www.google.com/patents/US20160179945>.
- [66] Lastra-Díaz, J. J., García-Serrano, A., 2016. WNSimRep: a framework and replication dataset for ontology-based semantic similarity measures and information content models. Mendeley Data v1, <http://dx.doi.org/10.17632/mpr2m8pycs.1>.
- [67] Lastra-Díaz, J. J., García-Serrano, A., 2016. WordNet-based word similarity reproducible experiments based on HESML V1R1 and ReproZip. Mendeley Data, v1, <http://dx.doi.org/10.17632/65pxgskhz9.1>.
- [68] Lastra-Díaz, J. J., García-Serrano, A., Batet, M., Fernández, M., Chirigati, F., Jun. 2017. HESML: a scalable ontology-based semantic similarity measures library with a set of reproducible experiments and a replication dataset. *Information Systems* 66, 97–118.
- [69] Laurence, S., Margolis, E., 1999. Concepts and Cognitive Science. In: Margolis, E., Laurence, S. (Eds.), *Concepts: Core readings*. MIT Press Cambridge, MA, pp. 3–81.

- [70] Leacock, C., Chodorow, M., 1998. Combining local context and WordNet similarity for word sense identification. In: Fellbaum, C. (Ed.), *WordNet: An electronic lexical database*. MIT Press, Ch. 11, pp. 265–283.
- [71] Lee, M. C., May 2011. A novel sentence similarity measure for semantic-based expert systems. *Expert Systems with Applications* 38 (5), 6392–6399.
- [72] Lee, M. D., Navarro, D. J., Nikkerud, H., 2005. An empirical evaluation of models of text document similarity. In: *Proceedings of the Cognitive Science Society*. Vol. 27. escholarship.org, pp. 1254–1259.
- [73] Lewis, D. D., Yang, Y., Rose, T. G., Li, F., 2004. RCV1: A New Benchmark Collection for Text Categorization Research. *Journal of machine learning research: JMLR* 5, 361–397.
- [74] Li, P., Wang, H., Zhu, K. Q., Wang, Z., Hu, X.-G., Wu, X., 2015. A Large Probabilistic Semantic Network based Approach to Compute Term Similarity. *IEEE Transactions on Knowledge and Data Engineering* 27 (10), 2604–2617.
- [75] Li, Y., Bandar, Z. A., McLean, D., 2003. An approach for measuring semantic similarity between words using multiple information sources. *IEEE Transactions on Knowledge and Data Engineering* 15 (4), 871–882.
- [76] Lidl, R., Pilz, G., 1998. *Applied Abstract Algebra*, 2nd Edition. Springer-Verlag, New York.
- [77] Lin, A. D., Graydon, P. J., Busch, J. E., Caudill, M., Chinchor, N. A., Tseng, J. C.-M., Wang, L., Pancho, B. S., Klein, K. S., Tijerino, Y. A., 6 Jan. 2004. Concept-based search and retrieval system. United States Patent and Trademark Office (USPTO) US 6675159 B1, <https://www.google.com/patents/US8301633>.
- [78] Lin, D., 1998. An information-theoretic definition of similarity. In: *Proceedings of the 15th International Conference on Machine Learning*. Vol. 98. Madison, WI, pp. 296–304.
- [79] Lin, Z., Kahng, M., Sabrin, K. M., Chau, D. H. P., Lee, H., Kang, U., Oct. 2014. MMap: Fast Billion-Scale Graph Computation on a PC via Memory Mapping. *Proceedings : ... IEEE International Conference on Big Data*. *IEEE International Conference on Big Data 2014*, 159–164.
- [80] Lord, P. W., Stevens, R. D., Brass, A., Goble, C. A., 1 Jul. 2003. Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics* 19 (10), 1275–1283.
- [81] Malewicz, G., Austern, M. H., Bik, A. J. C., Dehnert, J. C., Horn, I., Leiser, N., Czajkowski, G., 2010. Pregel: A System for Large-scale Graph Processing. In: *Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data*. *SIGMOD '10*. ACM, New York, NY, USA, pp. 135–146.

- [82] Mandreoli, F., Martoglia, R., Apr. 2011. Knowledge-based sense disambiguation (almost) for all structures. *Information Systems* 36 (2), 406–430.
- [83] Margolis, E., Jan. 1994. A reassessment of the shift from the classical theory of concepts to prototype theory. *Cognition* 51 (1), 73–89.
- [84] Margolis, E., Laurence, S. (Eds.), 1999. *Concepts: core readings*. MIT Press, Cambridge, MA.
- [85] Margolis, E., Laurence, S. (Eds.), 2015. *The Conceptual Mind: new directions in the study of concepts*. MIT Press, Cambridge, MA.
- [86] Martínez, S., Sánchez, D., Valls, A., 27 Oct. 2010. Ontology-Based Anonymization of Categorical Values. In: *Modeling Decisions for Artificial Intelligence*. Vol. 6408 of LNCS. Springer Berlin Heidelberg, pp. 243–254.
- [87] Martinez-Gil, J., Dec. 2016. CoTO: A Novel Approach for Fuzzy Aggregation of Semantic Similarity Measures. *Cognitive Systems Research* 40, 8–17.
- [88] Mazandu, G. K., Chimusa, E. R., Mulder, N. J., 29 Jul. 2016. Gene Ontology semantic similarity tools: survey on features and challenges for biological knowledge discovery. *Briefings in bioinformatics* [.http://dx.doi.org/10.1093/bib/bbw067](http://dx.doi.org/10.1093/bib/bbw067).
- [89] Meng, L., Gu, J., 2012. A New Model for Measuring Word Sense Similarity in WordNet. In: *Proceedings of the 4th International Conference on Advanced Communication and Networking, ASTL*. Vol. 14. pp. 18–23.
- [90] Meng, L., Gu, J., Zhou, Z., Sep. 2012. A new model of information content based on concept’s topology for measuring semantic similarity in WordNet. *International Journal of Grid and Distributed Computing* 5 (3), 81–93.
- [91] Meng, L., Huang, R., Gu, J., Jun. 2014. Measuring Semantic Similarity of Word Pairs Using Path and Information Content. *International Journal of Future Generation Communication & Networking* 7 (3), 183–194.
- [92] Metzler, D. A., Sep. 2007. Beyond bags of words: effectively modeling dependence and features in Information Retrieval. Ph.D. thesis, University of Massachusetts Amherst.
- [93] Meymandpour, R., Davis, J. G., Oct. 2016. A Semantic Similarity Measure for Linked Data: An Information Content-Based Approach. *Knowledge-Based Systems* 109, 276–293.
- [94] Mihalcea, R., Corley, C., Strapparava, C., 2006. Corpus-based and knowledge-based measures of text semantic similarity. In: *Proc. of the AAAI Conference on Artificial Intelligence*. Vol. 1. AAAI Press, pp. 775–780.
- [95] Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J., 2013. Distributed representations of words and phrases and their compositionality. In: Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z., Weinberger, K. Q. (Eds.),

- Advances in Neural Information Processing Systems 26 (NIPS 2013). NIPS Foundation, Inc., pp. 3111–3119.
- [96] Miller, G. A., Charles, W. G., 1991. Contextual correlates of semantic similarity. *Language and cognitive processes* 6 (1), 1–28.
- [97] Mohammad, S. M., Hirst, G., 8 Mar. 2012. Distributional Measures of Semantic Distance: A Survey. arXiv:1203.1858.
- [98] Monjardet, B., 1981. Metrics on partially ordered sets: A survey. *Discrete mathematics* 35 (1-3), 173–184.
- [99] Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., du Sert, N. P., Simonsohn, U., Wagenmakers, E.-J., Ware, J. J., Ioannidis, J. P. A., 10 Jan. 2017. A manifesto for reproducible science. *Nature Human Behaviour* 1, 0021.
- [100] Mustafa, J., Khan, S., Latif, K., 2008. Ontology based semantic information retrieval. In: *Intelligent Systems, 2008. IS'08. 4th International IEEE Conference*. Vol. 3. IEEE, Varna, pp. 22–14–22–19.
- [101] Ni, Y., Xu, Q. K., Cao, F., Mass, Y., Sheinwald, D., Zhu, H. J., Cao, S. S., 2016. Semantic Documents Relatedness Using Concept Graph Representation. In: *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining. WSDM '16*. ACM, New York, NY, USA, pp. 635–644.
- [102] Nosofsky, R. M., Mar. 1986. Attention, similarity, and the identification-categorization relationship. *Journal of experimental psychology. General* 115 (1), 39–57.
- [103] Nosofsky, R. M., 2011. The generalized context model: An exemplar model of classification. In: Pothos, E. M., Wills, A. J. (Eds.), *Formal Approaches in Categorization*. Cambridge University Press Cambridge, UK, Ch. 2, pp. 18–39.
- [104] Oliva, J., Serrano, J. I., del Castillo, M. D., Iglesias, A., Apr. 2011. SyMSS: A syntax-based measure for short-text semantic similarity. *Data & Knowledge Engineering* 70 (4), 390–405.
- [105] Orum, C., Joslyn, C. A., Mar. 2009. Valuations and Metrics on Partially Ordered Sets. arXiv:0903.2679.
- [106] Patwardhan, S., Banerjee, S., Pedersen, T., Feb. 2003. Using Measures of Semantic Relatedness for Word Sense Disambiguation. In: Gelbukh, A. (Ed.), *Proc. of the 4th International Conference on Computational Linguistics and Intelligent Text Processing (CICLING 2003)*. Vol. 2588 of LNCS. Springer, Mexico D.F., pp. 241–257.
- [107] Patwardhan, S., Pedersen, T., 2006. Using WordNet-based context vectors to estimate the semantic relatedness of concepts. In: *Proceedings of the EACL 2006 Workshop Making Sense of Sense-Bringing Computational Linguistics and Psycholinguistics Together*. Vol. 1501. Trento, Italy, pp. 1–8.

- [108] Pedersen, T., 2008. WordNet-InfoContent-3.0.tar dataset repository. https://www.researchgate.net/publication/273885902_WordNet-infoContent-3.0.tar.
- [109] Pedersen, T., 2010. Information Content Measures of Semantic Similarity Perform Better Without Sense-tagged Text. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. HLT '10. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 329–332.
- [110] Pedersen, T., Pakhomov, S. V. S., Patwardhan, S., Chute, C. G., Jun. 2007. Measures of semantic similarity and relatedness in the biomedical domain. *Journal of biomedical informatics* 40 (3), 288–299.
- [111] Pennington, J., Socher, R., Manning, C. D., 2014. Glove: Global vectors for word representation. *Proceedings of the Empirical Methods in Natural Language Processing (EMNLP 2014)* 12, 1532–1543.
- [112] Peraita Adrados, H., Labra, M. J., 1995. La obra de Eleanor Rosch veinte años después: Eleanor Rosch's work after twenty years. *Cognitiva* 7 (1), 67–92.
- [113] Pesquita, C., Faria, D., Falcao, A. O., Lord, P., Couto, F. M., 2009. Semantic similarity in biomedical ontologies. *PLoS computational biology* 5 (7), e1000443.
- [114] Pirró, G., Nov. 2009. A semantic similarity metric combining features and intrinsic information content. *Data & Knowledge Engineering* 68 (11), 1289–1308.
- [115] Pirró, G., Euzenat, J., 7 Nov. 2010. A Feature and Information Theoretic Framework for Semantic Similarity and Relatedness. In: Patel-Schneider, P. F., Pan, Y., Hitzler, P., Mika, P., Zhang, L., Pan, J. Z., Horrocks, I., Glimm, B. (Eds.), *Proc. of the 9th International Semantic Web Conference, ISWC 2010*. Vol. 6496 of LNCS. Springer, Shangai, China, pp. 615–630.
- [116] Pothos, E. M., Barque-Duran, A., Yearsley, J. M., Trueblood, J. S., Busemeyer, J. R., Hampton, J. A., 25 Feb. 2015. Progress and current challenges with the quantum similarity model. *Frontiers in psychology* 6, 205.
- [117] Pothos, E. M., Busemeyer, J. R., Trueblood, J. S., Jul. 2013. A quantum geometric model of similarity. *Psychological review* 120 (3), 679–696.
- [118] Pothos, E. M., Trueblood, J. S., Feb. 2015. Structured representations in a quantum probability model of similarity. *Journal of Mathematical Psychology* 64–65 (0), 35–43.
- [119] Rada, R., Mili, H., Bicknell, E., Blettner, M., 1989. Development and application of a metric on semantic nets. *IEEE Transactions on Systems, Man, and Cybernetics* 19 (1), 17–30.

- [120] Ramana Murty, P. V., Engelbert, T., 1985. On valuation in semilattices. *Mathematics Science Humanities* 90, 19–44.
- [121] Resnik, P., 20 Aug. 1995. Using information content to evaluate semantic similarity in a taxonomy. In: *Proc. of the International Joint Conferences on Artificial Intelligence (IJCAI 1995)*. Vol. 1. Montreal, Canada, pp. 448–453.
- [122] Rosch, E., 1999. Principles of categorization. *Concepts: Core Readings*, 189–206.
- [123] Rosch, E., Mervis, C. B., Oct. 1975. Family resemblances: Studies in the internal structure of categories. *Cognitive psychology* 7 (4), 573–605.
- [124] Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., Boyes-Braem, P., Jul. 1976. Basic objects in natural categories. *Cognitive psychology* 8 (3), 382–439.
- [125] Rubenstein, H., Goodenough, J. B., Oct. 1965. Contextual Correlates of Synonymy. *Communications of the ACM* 8 (10), 627–633.
- [126] Salton, G., Wong, A., Yang, C. S., Nov. 1975. A vector space model for automatic indexing. *Communications of the ACM* 18 (11), 613–620.
- [127] Sánchez, D., Batet, M., Oct. 2011. Semantic similarity estimation in the biomedical domain: an ontology-based information-theoretic perspective. *Journal of biomedical informatics* 44 (5), 749–759.
- [128] Sánchez, D., Batet, M., 2012. A new model to compute the information content of concepts from taxonomic knowledge. *International Journal on Semantic Web and Information Systems (ISWIS)* 8 (2), 34–50.
- [129] Sánchez, D., Batet, M., Isern, D., Mar. 2011. Ontology-based information content computation. *Knowledge-Based Systems* 24 (2), 297–303.
- [130] Sánchez, D., Batet, M., Isern, D., Valls, A., Jul. 2012. Ontology-based semantic similarity: A new feature-based approach. *Expert Systems with Applications* 39 (9), 7718–7728.
- [131] Schlicker, A., Lengauer, T., Albrecht, M., 15 Sep. 2010. Improving disease gene prioritization using the semantic similarity of Gene Ontology terms. *Bioinformatics* 26 (18), i561–7.
- [132] Sebastiani, F., 2002. Machine learning in automated text categorization. *ACM Computing Surveys (CSUR)* 34 (1), 1–47.
- [133] Sebt, A., Barfroush, A. A., Oct. 2008. A new word sense similarity measure in WordNet. In: *Proc. of the International Multiconference on Computer Science and Information Technology, IMCSIT 2008*. IEEE, pp. 369–373.
- [134] Seco, N., Veale, T., Hayes, J., 2004. An intrinsic information content metric for semantic similarity in WordNet. In: López de Mántaras, R., Saitta, L. (Eds.), *Proceedings of the 16th European Conference on Artificial Intelligence (ECAI)*. Vol. 16. IOS Press, Valencia, Spain, pp. 1089–1094.

- [135] Smith, E. E., Medin, D. L., 1999. The exemplar view. In: Margolis, E., Laurence, S. (Eds.), *Concepts core readings*. MIT Press, Cambridge, MA, pp. 207–221.
- [136] Stanchev, L., 2 Jun. 2014. Creating a Similarity Graph from WordNet. In: *Proceedings of the 4th International Conference on Web Intelligence, Mining and Semantics (WIMS'14)*. Article No. 36. ACM.
- [137] Turney, P. D., Pantel, P., 2010. From frequency to meaning: Vector space models of semantics. *The journal of artificial intelligence research* 37, 141–188.
- [138] Tversky, A., Jul. 1977. Features of similarity. *Psychological Review* 84 (4), 327–352.
- [139] Uren, V., Sabou, M., Motta, E., Fernandez, M., Lopez, V., Lei, Y., 1 Jan. 2010. Reflections on five years of evaluating semantic search systems. *International Journal of Metadata, Semantics and Ontologies* 5 (2), 87–98.
- [140] Vallet, D., Fernández, M., Castells, P., 2005. An ontology-based information retrieval model. In: *The Semantic Web: Research and Applications 2nd European Semantic Web Conference (ESWC 2005)*. Springer, Heraklion, Crete, Greece, pp. 455–470.
- [141] Vrandečić, D., Krötzsch, M., 23 Sep. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM* 57 (10), 78–85.
- [142] Wu, X., Pang, E., Lin, K., Pei, Z.-M., 31 May 2013. Improving the measurement of semantic similarity between gene ontology terms and gene products: insights from an edge- and IC-based hybrid method. *PloS one* 8 (5), e66745.
- [143] Wu, Z., Palmer, M., 1994. Verbs Semantics and Lexical Selection. In: *Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics. ACL '94*. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 133–138.
- [144] Yuan, Q., Yu, Z., Wang, K., Dec. 2013. A New Model of Information Content for Measuring the Semantic Similarity between Concepts. In: *Proc. of the International Conference on Cloud Computing and Big Data (CloudCom-Asia 2013)*. IEEE Computer Society, pp. 141–146.
- [145] Zhang, S.-B., Lai, J.-H., 11 Apr. 2016. Exploring information from the topology beneath the Gene Ontology terms to improve semantic similarity measures. *Gene* 586 (1), 148–157.
- [146] Zhou, Z., Wang, Y., Gu, J., 2008. A new model of information content for semantic similarity in WordNet. In: *Proc. of the Second International Conference on Future Generation Communication and Networking Symposia (FGCNS'08)*. Vol. 3. IEEE, pp. 85–89.

- [147] Zhou, Z., Wang, Y., Gu, J., Nov. 2008. New model of semantic similarity measuring in WordNet. In: Proc. of the 3rd International Conference on Intelligent System and Knowledge Engineering (ISKE 2008). Vol. 1. IEEE, pp. 256–261.

Part II

Publications and Patents

Chapter 7

Engineering Applications of Artificial Intelligence article

This page intentionally left blank.



ELSEVIER

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai

A novel family of IC-based similarity measures with a detailed experimental survey on WordNet

Juan J. Lastra-Díaz*, Ana García-Serrano

NLP & IR Research Group, ETSI Informática – UNED, Universidad Nacional de Educación a Distancia, C/Juan del Rosal 16, 28040 Madrid, Spain

ARTICLE INFO

Article history:

Received 13 April 2015

Received in revised form

27 August 2015

Accepted 6 September 2015

Keywords:

Ontology-based semantic similarity measures

IC-based measures

Semantic similarity

Intrinsic and corpus-based information content models

Jiang–Conrath distance

Semantic similarity on WordNet survey

ABSTRACT

This paper introduces a novel family of ontology-based similarity measures based on the Information Content (IC) theory, a detailed state of the art, a large experimental survey into ontology-based similarity measures on WordNet, and a new comparison between intrinsic and corpus-based IC models. Our experiments are based on our implementation of a large set of similarity measures, intrinsic and corpus-based IC models, which are evaluated on two known datasets and three different WordNet versions. The new measures are called *weighted Jiang–Conrath distance* ($wj&Cdist$) and *similarity* ($wj&Csim$), *cosine-normalized Jiang–Conrath similarity* ($cosj&Csim$) and *cosine-normalized weighted Jiang–Conrath similarity* ($coswj&Csim$). Two of our similarity measures outperform the state-of-the-art measures on the RG65 dataset, and one of them obtains the third overall score on all the datasets and evaluated WordNet versions. The cosine-normalized similarity measures are a non-linear normalization of the classic Jiang–Conrath (J&C) distance and the new $wj&C$ distance. On the other hand, the $wj&C$ distance is a generalization of the classic J&C distance which is based on the length of the shortest path between concepts within an IC-based weighted graph. Our measures are based on two not previously considered notions: (1) a generalization of the classic J&C distance to any type of taxonomy, based on an IC-based weighted graph derived from the conditional probabilities between child and parent concepts, and (2) a non-linear normalization function that converts the ontology-based semantic distances into similarity functions. Finally, the corpus-based IC models based on the Resnik method obtain rivaling results as regards the state-of-the-art intrinsic IC models, when they are used with some unexplored WordNet-based frequency files. Therefore, this latter fact allows us to reconsider some previous conclusions about the outperformance of the intrinsic IC models over the corpus-based ones.

© 2015 Elsevier Ltd. All rights reserved.

1. Introduction and positioning

The ontology-based similarity measures have found many applications in natural language processing (NLP), information retrieval (IR), and bioengineering. For example, in IR the aim is to retrieve resources that are semantically related to a user query both defined as concept sets. In this context, the word-to-word similarity measures can be extended to compute the distance between bags of concepts, or weighted concepts and individuals, thus, they are a key component in estimating the closeness between a user query and the relevant info to be retrieved. This approach is followed in Lastra-Díaz (2014), where we introduce a novel ontology-based IR model called *Intrinsic Ontological Spaces*, which is based on a metric space defined by the $wj&Cdistance$

* Corresponding author.

E-mail addresses: jlastra@invi.uned.es (J.J. Lastra-Díaz), agarcia@lsi.uned.es (A. García-Serrano).

introduced herein and disclosed in Lastra-Díaz and García-Serrano (2014). In Chan et al. (2011), the authors introduce a concept-based IR model for biomedical documents based on an ontology-based vector model, in which the document weights are computed using a non-linear function of a truncated version of the length of the shortest path between concepts. Next, we describe other applications of the ontology-based similarity measures. In Sánchez et al. (2015), the authors introduce the notion of semantic variance (SV) as a means of evaluating the quality of any ontology, which is defined as the variance of the semantic distance function, as defined in Batet et al. (2011), between each concept and the root. In Yan et al. (2014), the authors introduce an ontology-based inventive problem solving method which is based on a short-text similarity measure derived from the measure in Lin (1998). In Patwardhan et al. (2003), the authors introduce a word sense disambiguation (WSD) method based on the distributional hypothesis and the use of ontology-based similarity measures to select the closest evocated concept between a disambiguated

word and its neighboring words. In Mihalcea et al. (2006), the authors propose a text similarity measure based on the combination of an IDF weighting scheme with any ontology-based similarity measure, which is evaluated in a paraphrase detection (PD) task. In Cross and Hu (2011), the authors review the use of semantic similarity measures on the ontology alignment (OA) problem and introduce a semantic alignment quality measure based on the difference between the similarity measure between the concepts in the base ontology and their image in the target ontology. In Fiorini et al. (2015), the authors propose a semantic indexing method for biomedical documents based on similarity measures. In Couto and Pinto (2013) and Pesquita et al. (2009), the authors survey other applications of ontology-based similarity measures in bioengineering, such as the prediction of protein functions.

A semantic similarity measure is a binary function that given two input words computes their degree of similarity as perceived by a human being. Unlike the semantic relatedness between words, which includes other semantic co-occurrence relationships such as “part-of” or selectional preferences, the similarity measures are constrained to “is-a” relationships. The similarity measures can be roughly categorized into two families: ontology-based and corpus-based. An ontology-based semantic similarity measure is a binary function $sim : C \times C \rightarrow \mathbb{R}$ that approximates as much as possible the degree of similarity as perceived by a human being. In the latter expression, C is a concept set belonging to a taxonomy $\mathcal{C} = (C, \leq_C, \Gamma)$, which is defined by a partially ordered set (C, \leq_C) and an overall supreme element $\Gamma \in C$ called the root. A word is represented by a set of concepts within a base ontology, and the similarity between words is defined as the highest similarity value of the Cartesian product between both concept sets. On the other hand, most of corpus-based similarity measures are based on the distributional hypothesis Harris (1981), which states that words in similar contexts tend to share similar meanings. Most of distributional measures define the word meanings as a function of their context and the type of co-occurrence relationships that needs to be captured. For example, the word contexts could be small n -gram windows, or larger contexts such as sentences, paragraphs or documents. Despite there being different methods to represent the word meanings (contexts), such as sets, vectors, probability distributions, and graph nodes, the most popular representations rely on vector space models (VSM) (Turney and Pantel, 2010, Section 2.2). For example, in Gabrilovich and Markovitch (2007) the authors introduce a semantic relatedness method to compute word and document similarity, called ESA, which represents the meaning of a word or text as a weighted vector of Wikipedia concepts (articles), whose weights are defined by the cosine score between the input text vector and each Wikipedia base vector.

1.1. Ontology-based similarity measures versus corpus-based

The main advantage of the ontology-based measures is that the logic relationships between concepts, especially the “is-a” relationships, are hand-coded within the ontologies. A second advantage of these measures is that they are defined by closed formulas that only require a taxonomy to be evaluated. Therefore, they can be easily implemented, although their computational cost depends on the size of the ontology and the complexity of the required algorithms. In contrast, a serious drawback of the ontology-based measures in open domain applications, like the Web, is their limited lexical coverage, and the cost of creating and updating wide coverage ontologies. On the other hand, the corpus-based measures mainly rely on the distributional hypothesis, and compute the degree of similarity using an indirect approach that relies on the statistical co-occurrence between word contexts. In

addition to the “is-a” relationships, co-occurring words can encode other types of semantic relationships. Therefore, the corpus-based measures “can confuse similarity with relatedness” (Li et al., 2015, Section 1). Moreover, “it is commonly considered that distributional measures can only be used to capture semantic relatedness” (Harispe et al., 2015, Section 2.5.2) and “they have traditionally performed poorly when compared to WordNet-based measures” (Mohammad and Hirst, 2012, p1) in the similarity assessment task. Another drawback of the corpus-based measures is that they are commonly based on a pipeline of NLP and IR algorithms. According to the complexity of the measure, it could require syntactic pattern extraction, POS tagging, WSD and further methods, as well as external services and resources, leading them to a high computational cost and replication complexity. In addition, the corpus-based measures exhibit the classic problems related to the corpus statistics, such as the difficulty in obtaining a well-balanced corpus for all words and their senses. On the other hand, the main advantage of the corpus-based measures is that they offer a broader lexical coverage. In summary, the ontology-based similarity measures are efficient, robust and easy to implement, whilst the corpus-based measures offer a broader lexical coverage.

The mainstream of the research in corpus-based similarity measures is the proposal for hybrid concept-based distributional measures, which integrate KBs or explicit “is-a” semantic networks to bridge the lack of well-defined semantic knowledge. For example, in Patwardhan and Pedersen (2006) the authors introduce a similarity and relatedness measure which relies on the gloss vector overlapping between the extended WordNet gloss vectors of two input concepts. In Mohammad and Hirst (2006) the authors propose a hybrid distributional measure which relies on the cosine function and the concept-based conditional probabilities for the words derived from the Roget's thesaurus. In Alvarez and Lim (2007), the authors introduce another hybrid distributional similarity measure that relies on the product of two taxonomical WordNet-based functions with a gloss overlapping factor, which are defined on a WordNet subgraph that includes “is-a” and “part-of” relationships. Finally, in Li et al. (2013) and Li et al. (2015), the authors introduce a hybrid distributional measure whose core idea is that the similarity computation relies on truly “is-a” relationships, unlike traditional corpus-based measures. The Li et al. (2015) measure is based on a general-purpose “is-a” semantic network derived from a large web-based corpus. The semantic network is defined by a set of triplets $(c, e, (P(e|c), P(c|e)))$, where c is a hypernym of e . The words are categorized as concepts or entities depending on their hypernym hyponym ratio. The context of the concepts is defined as a vector whose weights are the conditional probabilities $P(e|c)$ of their subsumed entities, whilst the entity vectors are defined in the opposite way. The similarity is defined as the cosine function between concept vectors, or entity vectors. The underlying idea of the Li et al. method is to use the overlap between the extension sets (subsumed entities) of the concepts as an estimation of their similarity.

1.2. Focusing on ontology-based similarity measures

The recent progress in concept-based distributional measures has proven that this approach offers a good tradeoff between precision and lexical coverage, especially for general-purpose domains such as the Web. However, these measures require the semantic annotation of a large corpus with a good coverage of all the concepts required, which is not always possible. In addition, some of these large corpuses could be not publicly available. On the other hand, we prove herein that the ontology-based similarity measures even exhibit some margin of improvement with a low computational cost that deserves to be studied. In addition, there

are some specialized applications, like the protein function prediction problem in bioengineering, in which it is essential to provide precise conceptual models which cannot easily be extracted from a large corpus, or it is not a practical solution in these fields. Finally, both lines of research are complementary and closely related, and they can mutually benefit from their respective advances. For example, the semantic network used in Li et al. (2015) could use any WordNet-based ontology-based similarity measure if its concepts were mapped onto synsets. For these reasons, we focus our research effort herein on the proposal of new ontology-based similarity measures, more specifically those in the family of IC-based measures.

The taxonomy-based similarity estimation is a very old problem, which has been researched since the nineteen seventies in different fields, ranging from cognitive psychology Tversky (1977), to information retrieval Rada et al. (1989). Most of works categorize the different ontology-based semantic measures into three families, although there are also hybrid approaches. The main families of ontology-based similarity measures are as follows: (1) edge-counting measures, whose pioneering work has been carried-out by Rada et al. (1989), (2) IC-based measures, whose main references are the classic works of Resnik (1995), Jiang and Conrath (1997) and Lin (1998), (3) the feature-based measures, whose pioneering work has been carried-out by Tversky (1977), Sánchez et al. (2012) being the most recent, and (4) other intrinsic measures, such as Li et al. (2003) and Hadj Taieb et al. (2014b).

The development of novel intrinsic IC-based similarity measures is divided into two sub-problems: (1) the proposal of new IC-based similarity measures, as in our work, and (2) the development of new intrinsic IC models. The main drawbacks of the corpus-based IC models are the difficulty in getting a well-balanced and disambiguated corpus, the strong dependence on the training corpus, and the effort required for building the corpus and estimating the IC models. This fact has given rise to the development of the family of intrinsic IC models pioneered in the work of Seco et al. (2004), whose core idea is the computation of IC values using only the information encoded in the same ontology.

The state of the art in ontology-based semantic similarities and distances is currently defined by the intrinsic IC-based measures, which are defined by the combination of one intrinsic IC model with any IC-based measure. This statement is endorsed by several WordNet-based benchmarks reported in the literature, such as Budanitsky and Hirst (2006), Sánchez et al. (2011), Pirró (2009, fig.10), Hadj Taieb et al. (2014b) and the results reported herein. In a work in the field of bioengineering by Garla and Brandt (2012), the authors also prove experimentally that the intrinsic IC models outperform the corpus-based ones. However, we prove herein that the margin of performance between the intrinsic and corpus-based IC models is much smaller than the research community thought.

1.3. Main motivation

In Jiang and Conrath (1997), the authors prove that their semantic distance is equivalent to the shortest path between concepts over a weighted graph derived from the taxonomy, in which the edge weights are the IC values of the conditional probabilities between child and parent concepts. Nevertheless, this relationship has not been explored before, as it is herein. One drawback of the classic J&C distance, proven in Orum and Joslyn (2009), is that it is only a metric on tree-like taxonomies. Our work in Lastra-Díaz (2014) introduces a new ontology-based IR model based on a semantic metric space, and encouraged by this research, we introduce the *weighted Jiang–Conrath distance* which bridges the gap related to the drawback of the aforementioned J&C

distance. On the other hand, we also observe an underlying assumption in the literature as regards the conversion of any ontology-based distance, such as the J&C distance, into a similarity measure. In most cases, the J&C distance is converted into a similarity function through a linear mapping, such as the $sim_{J\&C}$ function shown in Table 2, despite this relationship being unknown and probably non-linear.

1.4. Definition of the problem and contributions

The main aim of this paper is to introduce a novel family of ontology-based similarity measures based on the Information Content theory to bridge the three gaps described in the paragraph above. These measures are called the *weighted Jiang–Conrath distance* ($wJ\&Cdist$), the *weighted Jiang–Conrath similarity* ($wJ\&Csim$), the *cosine-normalized Jiang–Conrath similarity* ($cosJ\&Csim$) and the *cosine-normalized weighted Jiang–Conrath similarity* ($coswJ\&Csim$). The new measures are based on a generalization of the Jiang–Conrath distance to any type of taxonomy through a weighted graph defined by the IC value of the edge-based conditional probabilities, and a normalization function that defines a non-linear mapping between the ontology-based semantic distances and its similarity value. Our main hypothesis is that integration of these two unexplored notions in the definition of the IC-based measures should lead us to improving the performance of the current IC-based similarity measures.

A second aim of this work is to reconsider some previous conclusions on the outperformance of the intrinsic IC model over the corpus-based ones, which rely on the results in Patwardhan and Pedersen (2006), Pedersen (2010) and Pirró (2009). These latter works are the primary sources that prove that the state-of-the-art intrinsic IC models outperform the corpus-based ones in the WordNet-based similarity tasks. The comparison and conclusions in other subsequent works, such as Sánchez et al. (2012) and Yuan et al. (2013), rely on these primary sources. However, the outperformance of the intrinsic IC models on the corpus-based models will be reconsidered in the light of the results obtained herein. This research is encouraged by the recent outstanding results of Gao et al. (2015), which question the conclusions of the research community on the outperformance of the intrinsic IC models over the corpus-based ones. The results reported by Gao et al. for their new hybrid IC-based similarity measure are based on a corpus-based IC model, which contradict to some degree our intuition and the mainstream of research led by the intrinsic IC-based similarity measures. This fact led us to the evaluation of all the IC-based similarity measures with a family of corpus-based IC models derived from some unexplored WordNet-based frequency files in Pedersen (2008). We prove herein that these corpus-based IC models obtain rivaling results as regards the state-of-the-art intrinsic IC models. Therefore, this latter fact allows us to reconsider the previous conclusions on the outperformance of the intrinsic IC models over the corpus-based ones. Despite the intrinsic IC models slightly outperforming the corpus-based models, we prove that the margin of performance between them is much smaller than the research community first thought.

This work also has other significant aims as follows. First, we carried-out a large and up-to-date experimental survey for most of the similarity measures on WordNet, which is based on our own implementation of most IC models and similarity measures reported in the literature. Second, our experiments allow the replication and validation of some previous approaches, as well as warning about the irreproducibility of others. Third, we carried-out an experimental study on the influence of the WordNet version on the similarity measures. Fourth, we study the performance of the similarity measures on two versions of the RG65 dataset, the classic one and the recent replication carried-out in Pirró (2009).

Table 1

State-of-the-art non IC-based similarity measures evaluated in our experiments.

Measure	Non IC-based similarity measures definition
Li et al. (2003)	$sim_{Li_{s3}}(c_1, c_2) = e^{-\alpha \cdot len}, \alpha^* = 0.25$
Li et al. (2003)	$sim_{Li_{s4}}(c_1, c_2) = e^{-\alpha \cdot len} \frac{e^{\beta \cdot d} - e^{-\beta \cdot d}}{e^{\beta \cdot d} + e^{-\beta \cdot d}}, \alpha^* = 0.2, \beta^* = 0.6$ $d = depth(LCA(c_1, c_2))$
Sánchez et al. (2012)	$dis_{SerB}(c_1, c_2) = \log_2 \left(1 + \frac{ \phi(c_1) \setminus \phi(c_2) + \phi(c_2) \setminus \phi(c_1) }{ \phi(c_1) \setminus \phi(c_2) + \phi(c_2) \setminus \phi(c_1) + \phi(c_1) \cap \phi(c_2) } \right)$ $\phi(a) = \{c \in C a \leq c\}$
Hadj Taieb et al. (2014a, b)	$sim_{T_{aieb_1}}(c_1, c_2) = TermDepth(c_1, c_2) \times TermHypo(c_1, c_2)$ $TermDepth(c_1, c_2) = \frac{2 \times depth(c_1, c_2)}{depth(c_1) + depth(c_2)}$ $TermHypo(c_1, c_2) = \frac{2 \times Spec_{Hypo}(c_1, c_2)}{Spec_{Hypo}(c_1, c_2) + Spec_{Hypo}(c_1, c_2)}$ $Spec_{Hypo}(c_1, c_2) = 1 - \frac{\log(HypoValue(c))}{\log(HypoValue(root))}$ $HypoValue(c) = \sum_{c' \in HypoInc(c)} P(depth(c'))$ $P(depth(c')) = \frac{ \{c' \in C depth(c') = depth(c)\} }{ C }$ $depth(c) = \text{length of the longest ascending path } c \rightarrow \text{root}$ $HypoInc(c) = \{c' \in C c' \leq c\}$
Hadj Taieb et al. (2014a, b)	$sim_{T_{aieb_2}}(c_1, c_2) = TermDepth(c_1, c_2) - \Lambda(w_1, w_2) $ $\times TermHypo(c_1, c_2)$ $\Lambda(w_1, w_2) = \max\{ \text{Synset}(w_1) , \text{Synset}(w_2) \}$

Fifth, our experimental survey of intrinsic and corpus-based IC-based similarity measures on WordNet is probably the most complete and up to date, providing a broad view of the state of the art of the problem. In our experiments, we evaluate and compare 17 similarity measures, 6 intrinsic IC models, 8 corpus-based IC models, 3 WordNet versions, and two datasets with similarity human judgments.

Despite our work belonging to the family of IC-based similarity measures, the two weighted similarity measures can be categorized in the subfamily of hybrid IC-based measures, whose main feature is the integration within the IC-based models and the measuring of features based on the length of the shortest path between concepts. Among the works in this family of hybrid IC-based measures we can cite the pioneering work of Li et al. (2003), as well as the works of Zhou et al. (2008b), Meng et al. (2014), Lastra-Díaz (2014) and Gao et al. (2015). In addition, our normalized measures are closely related to the non-linear scaling of the Lin similarity introduced in Meng and Gu (2012), although in our case, the non-linear normalization is encouraged by the unknown relationship between ontology-based distances and similarity measures.

The *cosJ&C* and *coswJ&C* similarity measures outperform the state-of-the-art measures on the RG65 dataset, while *cosJ&C* obtains the third overall score in our experiments. The corpus-based IC models derived from the classic Resnik method and the Pedersen dataset obtain rivaling results as regards the state-of-the-art intrinsic IC models. This latter fact allows us to review some previous conclusions in the corpus-based versus intrinsic IC models debate. According to our experimental results, the similarity measures with the best overall performance are the measures proposed in Hadj Taieb et al. (2014b) and Meng and Gu (2012), together with the *cosJ&C* and *coswJ&C* measures. It is

interesting to note that our measures and the measure proposed in Meng and Gu (2012) are non-linear normalizations of the classic Jiang–Conrath distance and the Lin similarity measure. Therefore, we prove that a proper normalization of some classic IC-based measures is enough to outperform the state-of-the-art methods at a low computational cost. This last fact allows the use of any hybrid IC-based measure to be refuted, because their performance does not justify their high computational cost. We confirm that there are no significant differences in the performance of the similarity measures in different WordNet versions. Finally, this work alerts us to the problem of the reproducibility of similarity measures, a problem that is also highlighted in Fokkens et al. (2013).

The rest of the paper is structured as follows. In Section 2, we review the literature on ontology-based similarity measures and Information Content models. Section 3 introduces the new family of IC-based semantic measures. In Section 4, we describe the evaluation methodology and the results obtained. Section 5 is devoted to our discussion. Finally, we summarize our conclusions and future work.

2. Ontology-based similarity measures and IC models

The literature on ontology-based semantic similarity measures and distances is very extensive, thus, we only focus on the measures that are evaluated in this work. First, we review the non IC-based similarity measures. Next, the rest of the section is devoted to review the family of IC-based measures and models, in which our work is framed. For a broader and recent survey on semantic similarity measures, we refer the reader to the recent book of Harispe et al. (2015). Further surveys can be found in Saruladha

et al. (2010), and Sánchez et al. (2012), as well as other surveys focused in bioengineering, such as Pesquita et al. (2009), Cross et al. (2013), Gan et al. (2013), and Harispe et al. (2014).

Modern research in the area starts with the work in Rada et al. (1989). In this work, the authors propose to use the length of the shortest path between concepts as a measurement of distance. Their work opens up the family of *edge-counting* semantic measures, and introduces the main hypothesis underlying all the subsequent ontology-based semantic distances: the *conceptual distance as a metrics* hypothesis. This hypothesis states, following previous psychological studies, that the conceptual distance, or similarity, between concepts in a semantic network, is proportional to the length of the path that links them. The ideas of Rada et al. are followed by other works, such as Wu and Palmer (1994), Leacock and Chodorow (1998) and Hirst and St-Onge (1998), which also propose similarity measures based on features derived from the length of shortest path between concepts.

In Tversky (1977), the authors introduce the first feature-based semantic similarity measure, which is defined by a weighted variant for the complement of the symmetric difference between the feature set of two concepts. With a perspective from set theory, the meaning of the Tversky measure is clear and well-founded. However, the feature sets associated to each concept cannot be derived directly from an ontology, which is a serious drawback for its practical implementation. With the aim of bridging the gap in the Tversky measure, in Sánchez et al. (2012), the authors introduce a feature-based dissimilarity measure shown in Table 1, which is based on the use of the common ancestors between concepts as a measure of their degree of similarity. The core idea behind the Sánchez et al. measure is that the ratio of overlap between common ancestors could be interpreted as an estimation of the ratio of common features between concepts, according to the Tversky model.

In Hadj Taieb et al. (2014b), the authors introduce two similarity measures as shown in Table 1. Although these measures are not based on an IC model, they are inspired and closely related to

the Seco et al. IC model. The core idea behind their approach is a new method of evaluating the contribution of the hyponym set of each concept to the similarity function, which relies on the use of the depth distribution on the taxonomy as a measure of the concept probabilities.

2.1. Similarity measures based on information content

The main drawback of the measures in the edge-counting family, called the *uniform weighting* premise, is that they implicitly assume that every edge has the same relevance in the computation of the overall length of the path, without considering its depth or probability of occurrence. With the aim of bridging this gap, Resnik introduces a new semantic similarity based on an Information Content (IC) model in his pioneering work Resnik (1995). The basic hypothesis behind all the IC-based similarity measures is that the more abstract concepts should have a lower information content than the more specific ones, and the higher the conditional probability between any concept and its parent, the shorter its distance. The IC measure for every concept $c_i \in C$ is the negative logarithm of its occurrence probability $p(c_i)$, as defined in Eq. (1). Resnik defines the similarity measure between two concepts as the IC value of the most informative common ancestor (MICA), as shown in Table 2.

$$IC(c_i) = -\log_2(p(c_i)) \tag{1}$$

One drawback of the Resnik similarity measure is that it only considers the IC value of the lowest ancestor concept, not the information along the path between concepts. With the aim of bridging this gap, in Jiang and Conrath (1997) the authors introduce the IC-based semantic distance shown in Table 2, whilst Lin introduces in Lin (1998) the similarity measure shown in same table. The J&C distance considers the two paths linking the evaluated concepts with their lowest common ancestor, and its definition is closely related to the metrics on lattices. In Orum and Joslyn (2009) the authors have proven that the J&C distance is only

Table 2
State-of-the-art IC-based similarity measures evaluated herein.

Measure	IC-based similarity measures definition
Resnik (1995)	$sim_{Resnik}(c_1, c_2) = IC(MICA(c_1, c_2))$
J&C (1997)	$d_{J\&C}(c_1, c_2) = IC(c_1) + IC(c_2) - 2IC(MICA(c_1, c_2))$ $sim_{J\&C}(c_1, c_2) = 1 - \frac{1}{2}d_{J\&C}(c_1, c_2)$
Lin (1998)	$sim_{Lin}(c_1, c_2) = \frac{2IC(MICA(c_1, c_2))}{IC(c_1) + IC(c_2)}$
Li et al. (2003)	$sim_{Li_{s9}}(c_1, c_2) = sim_{Li_{s4}}(c_1, c_2) * \frac{e^{\lambda*IC} - e^{-\lambda*IC}}{e^{\lambda*IC} + e^{-\lambda*IC}}, \lambda^* = 0.4$ $IC = MICA(c_1, c_2)$, see $sim_{Li_{s4}}$
Zhou et al. (2008a, b)	in Table 1 $sim_{Zh}(c_1, c_2) = 1 - k * \left(\frac{\log(\text{len}(c_1, c_2) + 1)}{\log\left(2 * \max_{c \in T} \{\text{depth}(c)\} - 1\right)} \right) - \frac{1}{2}(1 - k)d_{J\&C}(c_1, c_2)$ $k^* = \frac{1}{2}$ by default
P&S (2008)	$sim_{P\&S}(c_1, c_2) = \begin{cases} 3IC(MICA(c_1, c_2)) - IC(c_1) - IC(c_2) & \text{if } c_1 \neq c_2 \\ 1 & \text{if } c_1 = c_2 \end{cases}$
Meng and Gu (2012)	$sim_{Meng}(c_1, c_2) = e^{sim_{Lin}(c_1, c_2)} - 1 = e^{2IC(MICA(c_1, c_2))/IC(c_1) + IC(c_2)} - 1$
Meng et al. (2014)	$sim_{Meng2014}(c_1, c_2) = sim_{Lin}(c_1, c_2) * \left(\frac{1 - e^{-k * \text{len}(c_1, c_2)}}{e^{-k * \text{len}(c_1, c_2)}} \right)$, $k^* = 0.08$
Gao et al. (2015)	$sim_{Gao}(c_1, c_2) = e^{-\alpha L(c_1, c_2)}$, $\alpha^* = 0.15$ and $\beta^* = 2.05$ $L(c_1, c_2) = wt(c_1, c_2) * \text{len}(c_1, c_2)$ $wt = \begin{cases} \left(\frac{1 + IC(MICA(c_1, c_2))}{IC(MICA(c_1, c_2))} \right)^\beta & , IC(MICA(c_1, c_2)) \geq 1 \\ 2^\beta & , 1 > IC(MICA(c_1, c_2)) \geq 0 \end{cases}$

a metric on tree-like taxonomies, gap that is bridged by the new measure called the *weighted Jiang–Conrath distance* ($wJ\&Cdist$).

In Li et al. (2003), the authors introduce a family of ten different parametric similarity measures whose core idea is the breaking down of the overall similarity function into a combination of functions, where each base function relies on a different taxonomical feature. The taxonomical features used are the length of the shortest path between concepts, the depth of the lowest common ancestor, and the IC value of the MICA concept, as defined by Resnik. The IC-based measures introduced in this work include the first supervised and hybrid IC-based similarity measures reported in the literature. Unlike previous works, and the research carried-out herein, they compute the shortest path length and depth including the "is-a" and "has-a" links, thus, their results are not directly comparable to other methods, with the exception of the work of Pirró and Euzenat (2010). The Li et al. measures are defined by a set of free parameters which are trained on the RG65 \MC28 complementary subset, and evaluated on the MC28 dataset. For the IC-based measures they use the corpus-based IC model introduced in Resnik (1999). The main drawback of these measures is the need to tune the free parameters, as well as their influence in the results, which makes their configuration and generalization difficult. For the sake of completeness in our experimental comparison herein, we have implemented and evaluated the best three measures reported in Li et al. (2003), which are shown in Tables 1 and 2. These measures have not been evaluated on the RG65 dataset before, therefore, we also include herein their evaluation for first time, as well as the evaluation of their best IC-based measure, called strategy 9, with most intrinsic IC models. The measure called strategy 9 is, to our knowledge, the first hybrid IC-based measure reported in the literature, according to the aforementioned definition of this family.

In Pirró and Seco (2008), the authors introduce an IC-based similarity measure based on a reformulation of the Tversky measure in terms of the information content theory, which you can see in Table 2. They obtain good results and find a good connection between the feature-based and IC-based theories of similarity. In Pirró (2009), the author extends his previous paper with that of Seco. In the latter work, Pirró proves that a set of similarity measures based on the Seco et al. intrinsic IC model outperforms the same measures based on a corpus-based IC model derived from the Resnik method and the Brown corpus.

In Zhou et al. (2008b), the authors introduce the IC-based similarity measure shown in Table 2. This measure is defined by a linear combination of the classic J&C distance and a normalized value of the shortest path length between concepts.

In Meng and Gu (2012), the authors introduce the IC-based semantic similarity measure shown in Table 2, which is a non-linear transformation of the classic Lin measure, and is closely related to our work. In another recent work Meng et al. (2014), the authors introduce another variant of the Lin measure that is shown in Table 2. In this case, the similarity measure is a hybrid measure that combines the Lin IC-based measure with a power factor based on the shortest path length between concepts.

In Gao et al. (2015), the authors introduce three new similarity measures, among which we have the hybrid IC-based measure shown in Table 2. This similarity measure combines the Resnik measure with the $sim_{Li,s3}$ similarity function based on the length of the shortest path between concepts. The authors omit the IC model used in their experiments, however in a series of personal communications, they clarify that they used a corpus-based IC model based on the Resnik method that is also used in Patwardhan and Pedersen (2006). This fact encourages our evaluation of the corpus-based IC models provided in Pedersen (2008). The similarity measure shown in Table 2, called *strategy 3*, is a reformulation of the measure of Li et al. (2003), where the length factor

integrates an IC-based weight. This work is encouraged by the problem of parameter tuning in Li et al. (2003), where up to three parameters are used, which do not generalize well to other IC models and datasets. However, the proposed measure also requires two tuning parameters as in the work of Li et al. We implemented and evaluated the Gao et al. measure with most of the state-of-the-art intrinsic IC models in the literature, and the corpus-based IC models introduced herein. We obtained lower correlation values than those reported by the authors for similar conditions. In order to replicate exactly their results, the use of the same IC model used in their experiments would be necessary.

2.2. Intrinsic and corpus-based IC models

All the IC-based similarity measures require an IC model to be computed. An IC model is a concept-valued function that assigns an information content value to each ontology node. The drawbacks already described for the corpus-based IC models have encouraged the development of intrinsic IC models, whose pioneering work is the intrinsic IC model of Seco et al. (2004). Other intrinsic IC models reported in the literature are the works in Zhou et al. (2008a), Sebtí and Barfroush (2008), Sánchez et al. (2011), Sánchez and Batet (2012), Yuan et al. (2013), Meng et al. (2012), and Hadj Taieb et al. (2014a). In Table 3, we show the intrinsic IC models that are evaluated in our experiments.

In Pedersen (2013), the author explains how to apply the Resnik method, introduced in Resnik (1999), to compute an IC model using his WordNet-based frequency files Pedersen (2008). Pedersen uses this method to build the corpus-based IC models in Patwardhan and Pedersen (2006) and Pedersen (2010). Pirró also confirms that a corpus-based IC model based on the Resnik method and the Brown Corpus is used in Pirró (2009). In order to make a comparison between intrinsic and corpus-based and IC models, we implemented the Resnik method to build some corpus-based IC models with some unexplored frequency files in the Pedersen dataset. In the light of the results obtained herein, we propose to use these corpus-based IC models as a new baseline for the evaluation of IC-based similarity measures and IC models.

The dataset in Pedersen (2008) includes a family of WordNet-based frequency files derived from the British National Corpus, the Brown corpus, the SemCor and SemCorRaw corpus, the Penn Treebank and the complete works of Shakespeare. Within this family of files, we find a subset with the suffix "add1". These files start the count for any concept to 1 to guarantee that there are no concepts with zero frequency. We have used these unexplored files to build all the corpus-based IC models herein, obtaining the higher correlation values reported in the literature for any corpus-based IC model evaluated on WordNet. These results rival the state-of-the-art intrinsic IC models, which led us to refute the previous conclusions founded in the experiments of Pedersen and Pirró.

$$f : C \rightarrow \mathbb{N} \quad (2)$$

$$f(c_i) = TF(c_i) + IF(c_i) = TF(c_i) + \sum_{\forall c_j | c_i \in LA(c_j)} f(c_j) \quad (3)$$

$$\hat{p}(c_i) = \frac{f(c_i)}{N} = \frac{f(c_i)}{f(I)} \quad (4)$$

In Resnik (1999), the author introduces the most broadly accepted corpus-based IC model for the evaluation of semantic similarity tasks. Resnik proposes a method to compute an IC model that is based on the estimation of the concept probabilities through the frequency counting of concept occurrences in a training corpus. Each occurrence in the corpus of a word contained in WordNet is counted as an occurrence of all its subsumer

Table 3
State-of-the-art intrinsic IC models evaluated in this work.

IC models	Definition
Seco et al. (2004)	$IC_{Seco}(c) = 1 - \frac{\log(Hypo(c) + 1)}{\log(max_nodes)}$
Zhou et al. (2008a, b)	$IC_{Zhou}(c) = k \left(1 - \frac{\log(Hypo(c) + 1)}{\log(max_nodes)} \right) + (1-k) \frac{\log(depth(c))}{\log(depth_{max})}$, $k^* = \frac{1}{2}$ (default)
Sánchez et al. (2011)	$IC_{Sánchez2011}(c) = -\log_2 \left(\frac{ Leaves(c) }{ Subsumers(c) + 1} \right)$
Sánchez et al. (2012)	$IC_{Sánchez2012}(c) = -\log_2 \left(\frac{commonness(c)}{commonness(root)} \right)$ $\begin{cases} commonness(c) = \frac{1}{ Subsumers(c) } & , c \text{ leaf} \\ commonness(c) = \sum commonness(l) & , c \text{ not leaf} \\ \forall l l \text{ is leaf and } l < c \end{cases}$
Meng et al. (2012)	$IC_{Meng}(c) = \frac{\log(depth(c))}{\log(depth_{max})} \times \left(1 - \frac{\log \left(1 + \sum_{a \in Hypo(c)} \frac{1}{\log(Node_{max})} \right)}{\log(Node_{max})} \right)$
Yuan et al. (2013)	$IC_{Yuan}(c) = f_{depth}(c)(1 - f_{leaves}(c)) + f_{hyper}(c)$ $\begin{cases} f_{depth}(c) = \frac{\log(depth(c))}{\log(depth_{max})} \\ f_{leaves}(c) = \frac{\log(Leaves(c) + 1)}{\log(Leaves_{max} + 1)} \\ f_{hyper}(c) = \frac{\log(Hyper(c) + 1)}{\log(Node_{max})} \end{cases}$

concepts. In Pedersen (2013, p. 34), the author describes the Resnik frequency counting method used to build the WordNet-based frequency files dataset, Pedersen (2008), as well as the corpus-based IC models evaluated in his paper series on similarity measures on WordNet. Following the notation of Pedersen, each concept frequency $f(c_i)$ in Eq. (2) is defined as the sum of the term-frequency (TF) occurrences of the concept c_i , plus the inherited frequency (IF) of each subsumed child concept. The estimated probability $\hat{p}(c_i)$ of each taxonomic concept $c_i \in C$ is defined in Eq. (4), where N is the total number of occurrences of any noun within the corpus and its value matches the frequency of the root concept Γ . Finally, the IC values are computed using Eq. (1). The frequency counting proposed by Resnik does not take into account the word senses, although Resnik suggest that a sense-tagged corpus could be used to improve this issue. In another work, Pedersen (2010), the authors prove that the IC models derived from a non-sense-tagged corpus perform better than the sense-tagged ones.

3. The new IC-based similarity measures

In this section, we introduce a new ontology-based distance, defined in Eq. (7), and three new IC-based similarity measures, defined in Eqs. (8), (11) and (12) below. These measures are based on two different unexplored notions: (1) a generalization of the Jiang–Conrath distance, and (2) a non-linear normalization for the conversion of ontology-based semantic distances into similarity measures.

The normalization function is based on the computation of the maximum distance on the taxonomy, and a cosine-based scaling function φ_c that provides an exponential-like mapping for modelling the distance-to-similarity transformation. Despite that herein we only consider some variants of the Jiang–Conrath distance; our normalization function could be used with other semantic distances. The normalization proposed herein is closely

related to the scaled variant of the Lin measure that is introduced in Meng and Gu (2012). However, our scaling function φ_c is different, and our main motivation is a first try at modelling the unknown relationships between ontology-based distances and similarity measures; a problem that has not been studied before.

Any taxonomy on a set of concepts C is defined formally by a $\mathcal{C} = (C, \leq_c, \Gamma)$ triplet, where (C, \leq_c) is a partially ordered set, and $\Gamma \in C$ is a distinguished element called the root concept, such that $\forall c_i \in C \rightarrow c_i \leq_c \Gamma$. Every taxonomy $\mathcal{C} = (C, \leq_c, \Gamma)$ induces a graph $G = (E, V)$ in the usual manner, where every concept $c_i \in C$ is a vertex of the graph, and there is an edge between each concept c_i and its direct parents, also called the lowest ancestors of c_i , and denoted as $LA(c_i)$.

$$d_{J\&C}(c_1, c_2) = IC(c_1) + IC(c_2) - 2IC(MICA(c_1, c_2)) \quad (5)$$

In Jiang and Conrath (1997), the authors prove that their semantic distance $d_{J\&C}(c_1, c_2)$, as defined in Eq. (5), is equivalent to the shortest path between concepts c_1 and c_2 on a weighted graph derived from the taxonomy, whose edge weights are defined in the following equation:

$$w : E \rightarrow \mathbb{R} \quad (6)$$

$$w(e_{ij}) = -\log_2(p(c_i | c_j)) = IC(c_i) - IC(c_j)$$

$$E = \{(c_i, c_j) \subset C \times C | c_j \in LA(c_i)\}$$

Despite Jiang and Conrath claiming that their distance is a metric on any type of taxonomy, in Orum and Joslyn (2009), the authors prove that it is only true for tree-like taxonomies. Indeed, Eq. (6) is only satisfied on tree-like taxonomies, discarding any taxonomy with a multiple inheritance, such as WordNet. In order to define a well-founded metric on any type of taxonomy, we propose the *weighted Jiang–Conrath distance*, as defined in Eq. (7), together with its similarity function $sim_{wJ\&C}(c_1, c_2)$ defined in Eq. (8). The $d_{wJ\&C}(c_1, c_2)$ is defined as the weighted shortest path between concepts, which can be computed by any known method, such as

the Dijkstra algorithm. The distance function $d_{w\&c}(c_1, c_2)$ matches exactly the classic Jiang–Conrath distance on a tree-like taxonomy regardless of the underlying IC model. This latter fact is a theorem that can be proven and it only depends on the structure on the taxonomy, not the IC model.

$$d_{w\&c}(c_1, c_2) = \min_{\forall \alpha \in \text{Paths}_{(c_1, c_2)}} \left\{ \sum_{e_{ij} \in \alpha} w(e_{ij}) \right\} \quad (7)$$

$$\text{sim}_{w\&c}(c_1, c_2) = 1 - \frac{d_{w\&c}(c_1, c_2)}{2} \quad (8)$$

The edge weights $w(e_{ij})$ above are defined in two different ways in Eq. (9), according to the availability of estimated values for the conditional probabilities $p(c_i|c_j)$. Whenever the $p(c_i|c_j)$ function is known, the weights are defined as the IC value of the conditional probabilities. Otherwise, the weights are defined as the absolute difference of node-based IC values, which allows the integration of any IC model reported in the literature.

$$w : E \rightarrow \mathbb{R} \quad (9)$$

$$w(e_{ij}) = \begin{cases} -\log_2(p(c_i|c_j)) & \text{if } p(c_i|c_j) \text{ are known} \\ |IC(c_i) - IC(c_j)| & \text{otherwise} \end{cases}$$

The new *weighted Jiang–Conrath distance* matches the classic Jiang–Conrath on tree-like taxonomies, however, it is slightly different on taxonomies with a multiple inheritance. The computation of the $d_{w\&c}$ function requires the implementation of any shortest path algorithm, our preferred method being a version of the classic Dijkstra algorithm based on a min-priority queue (Chen et al., 2007).

In order to model the unknown relationship between the ontology-based distances and similarities, we propose the φ_c function defined in the following equation:

$$\begin{aligned} \varphi_c : [0, 1] \subset \mathbb{R} &\rightarrow [0, 1] \subset \mathbb{R} \\ \varphi_c(x) &= 1 - \cos\left(\frac{\pi}{2}x\right) \end{aligned} \quad (10)$$

Finally, we propose two new normalized similarity measures as follows. First, the *cosine-normalized weighted J&C similarity*, which is derived from the new $d_{w\&c}(c_1, c_2)$ distance and is denoted by $\text{sim}_{\text{cos}w\&c}$ in Eq. (11). Second, the *cosine-normalized J&C similarity*, which is derived from the classic $d_{j\&c}$ distance and is denoted by $\text{sim}_{\text{cos}j\&c}$ in Eq. (12).

Both similarity functions are obtained through the composition of their respective base distance with the non-linear function φ_c . Before the distance-to-similarity conversion given by the φ_c function, both distances are normalized by the maximum distance between the root concept Γ and any leaf concept within the taxonomy, whose value is defined by $\max_{d_{j\&c}}$ in Eq. (13). By doing some algebra and recalling that $IC(\Gamma)$ is equal to 0, we can see in Eq. (14) that $\max_{d_{j\&c}}$ is equal to the maximum IC value within the set of leaf concepts.

$$\begin{aligned} \text{sim}_{\text{cos}w\&c}(c_1, c_2) &= \varphi_c \circ \left(1 - \frac{d_{w\&c}(c_1, c_2)}{2 * \max_{d_{j\&c}}} \right) \\ &= 1 - \cos\left(\frac{\pi}{2} \left(1 - \frac{d_{w\&c}(c_1, c_2)}{2 * \max_{d_{j\&c}}} \right)\right) \end{aligned} \quad (11)$$

$$\begin{aligned} \text{sim}_{\text{cos}j\&c}(c_1, c_2) &= \varphi_c \circ \left(1 - \frac{d_{j\&c}(c_1, c_2)}{2 * \max_{d_{j\&c}}} \right) \\ &= 1 - \cos\left(\frac{\pi}{2} \left(1 - \frac{d_{j\&c}(c_1, c_2)}{2 * \max_{d_{j\&c}}} \right)\right) \end{aligned} \quad (12)$$

$$\max_{d_{j\&c}} = \max_{c \in \text{Leaves}(C)} \{d_{j\&c}(\Gamma, c)\} \quad (13)$$

$$= \max_{c \in \text{Leaves}(C)} \{IC(c)\} \quad (14)$$

4. Evaluation

The goals of the experimental work described in this section are as follows: (1) the experimental evaluation and comparison of the new IC-based similarity measures with most of similarity measures, intrinsic and corpus-based IC models reported in the literature, (2) the replication of previously reported methods and results, (3) the experimental study into the influence of the WordNet version on the similarity measures, (4) a comparison between intrinsic and corpus-based IC models, (5) a study of the performance of the similarity measures on two versions of the RG65 dataset, (6) the refutation of some previous conclusions on the performance of corpus-based IC models versus the intrinsic IC models, and (7) a new confirmation of the achievements of the family of intrinsic IC models.

4.1. Similarity measures and IC-based models evaluated

In order to compare our new similarity measures defined in Eqs. (8), (11) and (12), with the state-of-the-art measures, we implemented all the similarity measures shown in Tables 1 and 2, as well as all the intrinsic IC models shown in Table 3. For all the IC models and similarity measures implemented herein, we consider the depth as the length of shortest ascending path from each concept to the root node, with the exception of the Hadj Taieb et al. measures, where they explicitly define it as the longest ascending path length. For the Zhou et al. IC model and measure, the authors define the depth starting at 1 for the root concept. Our primary aims lead us to implementing all methods evaluated herein, but the readers working on applications could use the SML library (Harispe et al., 2014).

In addition to the intrinsic IC models in Table 3, we also implemented a set of corpus-based IC models based on the family of “add1” WordNet-based frequency files of Pedersen (2008) and the Resnik method aforementioned. We show the comparison results in Tables 6 and 7, which is the largest experimental survey between intrinsic and corpus-based IC models reported in the literature.

4.2. Experimental setup

For the experiments, we use the noun database of WordNet 2.1, 3.0 and 3.1 versions Miller (1995), together with the classic RG65 dataset introduced in Rubenstein and Goodenough (1965), and a recent replication of it, called $P\&S_{full}$ and introduced in Pirró (2009). The RG65 dataset is made up of 65 word pairs together with their similarity human judgments in the range (0, 4). In order to manage the polysemy in WordNet, we evaluate the similarity measure for the Cartesian product of the synsets of the input word pair, and we choose the higher similarity result, as done in other works in the field, such as Sánchez et al. (2012, Section 3.4). As an evaluation metric, we use the standard Pearson correlation factor, as defined in the following equation:

$$r = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}} \quad (15)$$

Table 4Best Pearson correlation value for all the similarity measures evaluated on the RG65 and $P\&S_{full}$ datasets, using different versions of WordNet.

WordNet versions Measures/Datasets	WordNet 2.1		WordNet 3.0		WordNet 3.1		Overall score (avg. correlation)		
	RG65	$P\&S_{full}$	RG65	$P\&S_{full}$	RG65	$P\&S_{full}$	RG65	$P\&S_{full}$	Overall
Taieb sim1 (2014)	0.8673	0.9074	0.8670	0.9068	0.8670	0.9067	0.8671	0.9070	0.8870
Meng and Gu (2012)	0.8680	0.9067	0.8675	0.9061	0.8675	0.9061	0.8677	0.9063	0.8870
cosj&C (this work)	0.8701	0.8942	0.8752	0.8996	0.8751	0.8996	0.8735	0.8978	0.8856
Taieb sim2 (2014)	0.8614	0.9066	0.8597	0.9050	0.8600	0.9052	0.8604	0.9056	0.8830
Zhou et al. (2008a, b)	0.8681	0.8905	0.8728	0.8949	0.8708	0.8940	0.8706	0.8931	0.8818
coswj&C (this work)	0.8763	0.8843	0.8770 (this work)	0.8875	0.8746	0.8843	0.8760	0.8854	0.8807
Gao et al. (2015)	0.8676	0.8926	0.8709	0.8920	0.8682	0.8919	0.8689	0.8922	0.8805
Lin (1998)	0.8638	0.8915	0.8621	0.8948	0.8621	0.8948	0.8627	0.8937	0.8782
Pirró and Seco (2008)	0.8596	0.8903	0.8622	0.8970	0.8622	0.8970	0.8613	0.8948	0.8780
Li et al. strat9 (2003)	0.8613	0.8895	0.8617	0.8897	0.8615	0.8897	0.8615	0.8896	0.8756
Li et al. strat3 (2003)	0.8633	0.8863	0.8625	0.8853	0.8594	0.8840	0.8617	0.8852	0.8735
Li et al. strat4 (2003)	0.8580	0.8779	0.8598	0.8787	0.8598	0.8787	0.8592	0.8784	0.8688
Jiang and Conrath (1997)	0.8566	0.8701	0.8619	0.8825	0.8619	0.8781	0.8601	0.8769	0.8685
Resnik (1995)	0.8418	0.8826	0.8409	0.8829	0.8409	0.8829	0.8412	0.8828	0.8620
Sánchez et al. (2012)	0.8545	0.8780	0.8477	0.8703	0.8477	0.8703	0.8500	0.8729	0.8614
wj&C (this work)	0.8555	0.8515	0.8618	0.8697	0.8606	0.8602	0.8593	0.8604	0.8599
Meng et al. (2014)	0.8463	0.8351	0.8486	0.8374	0.8486	0.8374	0.8478	0.8367	0.8423
Best column value	0.8763	0.9074	0.8770	0.9068	0.8751	0.9067	0.8760	0.9070	0.8870

Table 5

Correlation values for the IC-based similarity measures evaluated on the RG65 and Pirró datasets with WordNet 2.1.

Intrinsic IC models	State-of-the-art IC-based similarity measures									New measures (this work)		
	Pearson correlation values for the IC-based similarity measures on the RG65 dataset and WordNet 2.1											
	Resnik	Lin	J&C	Li_s9	Zhou	P&S	Meng12	Meng14	Gao	wj&C	cosj&C	coswj & C
Seco et al. (2004)	0.8323	0.8562	0.8456	0.8240	0.8681	0.8543	0.8564	0.8463	0.7992	0.8424	0.8568	0.8558
Zhou et al. (2008a, b)	0.8117	0.8293	0.8256	0.8429	0.8561	0.8316	0.8552	0.7737	0.7992	0.8296	0.8534	0.8562
Sánchez et al. (2011)	0.8418	0.8516	0.8522	0.8568	0.8548	0.8022	0.8648	0.8118	0.8676	0.8524	0.8681	0.8763
Sánchez et al. (2012)	0.8354	0.8573	0.8423	0.8613	0.8456	0.8232	0.8569	0.8452	0.8675	0.8418	0.8535	0.8544
Meng et al. (2012)	0.8289	0.8621	0.8566	0.8277	0.8645	0.8577	0.8671	0.8277	0.7992	0.8555	0.8701	0.8715
Yuan et al. (2013)	0.8274	0.8638	0.8479	0.8224	0.8625	0.8596	0.8680	0.8257	0.7992	0.8432	0.8615	0.8593
Best column value	0.8418	0.8638	0.8566	0.8613	0.8681	0.8596	0.8680	0.8463	0.8676	0.8555	0.8701	0.8763
Intrinsic IC models	Pearson correlation values for the IC-based measures on the $P\&S_{full}$ dataset and WordNet 2.1											
Seco et al. (2004)	0.8795	0.8908	0.8701	0.8839	0.8905	0.8903	0.9008	0.8351	0.8587	0.8515	0.8904	0.8753
Zhou et al. (2008a, b)	0.8401	0.8458	0.8350	0.8895	0.8712	0.8556	0.8823	0.7475	0.8587	0.8197	0.8708	0.8542
Sánchez et al. (2011)	0.8751	0.8731	0.8685	0.8799	0.8716	0.7988	0.8955	0.7914	0.8843	0.8475	0.8942	0.8843
Sánchez et al. (2012)	0.8826	0.8915	0.8667	0.8883	0.8701	0.8314	0.9012	0.8336	0.8926	0.8511	0.8859	0.8731
Meng et al. (2012)	0.8679	0.8879	0.8675	0.8813	0.8803	0.8884	0.9029	0.8101	0.8587	0.8495	0.8893	0.8738
Yuan et al. (2013)	0.8686	0.8915	0.8600	0.8769	0.8783	0.8875	0.9067	0.8069	0.8587	0.8437	0.8813	0.8675
Best column value	0.8826	0.8915	0.8701	0.8895	0.8905	0.8903	0.9067	0.8351	0.8926	0.8515	0.8942	0.8843

4.3. Results

In Table 4, we show a summary of the best correlation values obtained for the similarity measures evaluated in the RG65 and $P\&S_{full}$ datasets, using the latest three versions of WordNet. The overall score is the average value of the correlation values obtained by each measure in each combination of dataset and WordNet versions. The measures are sorted according to their score value shown in last column. The only difference between both datasets are the human judgments, however, we confirm that their joint correlation value is 0.972, as reported in Pirró (2009). In each column, the best correlation value is shown in bold. In Tables 5 and 8 we show the complete results for the evaluation of all the IC-based similarity measures in WordNet versions 2.1 and 3.1. In Table 6 we show the complete results for the IC-based similarity measures evaluated on the RG65 dataset and WordNet 3.0, including a set of corpus-based IC models, while in Table 7 we do the same in the $P\&S_{full}$ dataset.

5. Discussion

Our new IC-based similarity measures called *coswj&C* and *cosj&C* obtain the highest correlation values in the classic RG65 dataset and all the WordNet versions. The first intrinsic Hadj Taieb et al. similarity measure obtains the higher correlation values in the $P\&S_{full}$ dataset and all the WordNet versions. Finally, the Hadj Taieb et al. measure and Meng et al. (2012) IC-based similarity measure obtain the best overall scores when the correlation values are averaged over all the datasets and WordNet versions.

These results confirm that the state of the art on the problem is lead by the family of IC-based similarity measures on WordNet, although all the IC-based similarity measures show a strong dependence between their performance and the IC models used. Two of our novel measures and the Meng et al. (2012) measure are non-linear normalizations of the classic IC-based similarity measures introduced by Lin and Jiang–Conrath, whilst the intrinsic similarity measure introduced by Hadj taieb et al. is inspired by

Table 6

Correlation values for the IC-based similarity measures evaluated on the RG65 dataset with WordNet 3.0, using intrinsic and corpus-based IC models.

Intrinsic IC models	State-of-the-art IC-based similarity measures								New measures (this work)			
	Pearson correlation values for the IC-based similarity measures on the RG65 dataset and WordNet 3.0											
	Resnik	Lin	J&C	Li_s9	Zhou	P&S	Meng12	Meng14	Gao	wj&C	cosj&C	coswj&C
Seco et al. (2004)	0.8326	0.8609	0.8546	0.8241	0.8728	0.8622	0.8596	0.8486	0.7992	0.8520	0.8642	0.8634
Zhou et al. (2008a, b)	0.8080	0.8259	0.8286	0.8438	0.8574	0.8334	0.8539	0.7749	0.7992	0.8337	0.8558	0.8610
Sánchez et al. (2011)	0.8409	0.8530	0.8619	0.8586	0.8639	0.8105	0.8663	0.8147	0.8682	0.8526	0.8752	0.8770
Sánchez et al. (2012)	0.8355	0.8616	0.8508	0.8615	0.8535	0.8332	0.8600	0.8475	0.8678	0.8504	0.8606	0.8614
Meng et al. (2012)	0.8260	0.8608	0.8598	0.8282	0.8638	0.8586	0.8670	0.8285	0.7992	0.8613	0.8723	0.8747
Yuan et al. (2013)	0.8243	0.8621	0.8505	0.8231	0.8629	0.8607	0.8675	0.8273	0.7992	0.8474	0.8632	0.8624
Best column value	0.8409	0.8621	0.8619	0.8615	0.8728	0.8622	0.8675	0.8486	0.8682	0.8613	0.8752	0.8770
Corpus-based IC models	Pearson correlation values for the IC-based similarity measures on the RG65 dataset and WordNet 3.0											
ic-bnc-resnik-add1	0.8281	0.8543	0.8609	0.8603	0.8633	0.8311	0.8591	0.8377	0.8673	0.8587	0.8644	0.8644
ic-brown-resnik-add1	0.8293	0.8519	0.8531	0.8559	0.8559	0.8268	0.8555	0.8389	0.8659	0.8514	0.8581	0.8597
ic-semcor-add1	0.8257	0.8506	0.8551	0.8605	0.8577	0.8279	0.8522	0.8402	0.8660	0.8539	0.8590	0.8605
ic-semcorraw-add1	0.8363	0.8569	0.8595	0.8608	0.8620	0.8342	0.8581	0.8432	0.8680	0.8618	0.8658	0.8700
ic-semcorraw-resnik-add1	0.8345	0.8564	0.8560	0.8602	0.8590	0.8301	0.8581	0.8417	0.8665	0.8551	0.8622	0.8640
ic-shaks-resnik-add1	0.8223	0.8502	0.8528	0.8596	0.8551	0.8185	0.8549	0.8366	0.8640	0.8582	0.8638	0.8703
ic-treebank-add1	0.8345	0.8589	0.8561	0.8617	0.8582	0.8335	0.8609	0.8434	0.8709	0.8579	0.8653	0.8703
ic-treebank-resnik-add1	0.8331	0.8542	0.8536	0.8603	0.8564	0.8259	0.8589	0.8392	0.8673	0.8539	0.8618	0.8661
Best column value	0.8363	0.8589	0.8609	0.8617	0.8633	0.8342	0.8609	0.8434	0.8709	0.8618	0.8658	0.8703
Overall best value	0.8409	0.8621	0.8619	0.8617	0.8728	0.8622	0.8675	0.8486	0.8709	0.8618	0.8752	0.8770

Table 7

Correlation values for the IC-based similarity measures evaluated on the P&S_{full} dataset with WordNet 3.0, using intrinsic and corpus-based IC models.

Intrinsic IC models	State-of-the-art IC-based similarity measures								New measures (this work)			
	Pearson correlation values for the IC-based similarity measures on the P&S _{full} dataset and WordNet 3.0											
	Resnik	Lin	J&C	Li_s9	Zhou	P&S	Meng12	Meng14	Gao	wj&C	cosj&C	coswj&C
Seco et al. (2004)	0.8799	0.8945	0.8781	0.8839	0.8949	0.8970	0.9031	0.8374	0.8585	0.8601	0.8966	0.8819
Zhou et al. (2008a, b)	0.8357	0.8420	0.8372	0.8897	0.8725	0.8563	0.8806	0.7488	0.8585	0.8247	0.8726	0.8593
Sánchez et al. (2011)	0.8740	0.8738	0.8762	0.8807	0.8789	0.8051	0.8964	0.7943	0.8844	0.8483	0.8996	0.8850
Sánchez et al. (2012)	0.8829	0.8948	0.8742	0.8877	0.8771	0.8398	0.9035	0.8350	0.8920	0.8587	0.8918	0.8790
Meng et al. (2012)	0.8645	0.8863	0.8715	0.8813	0.8805	0.8897	0.9025	0.8106	0.8585	0.8550	0.8917	0.8765
Yuan et al. (2013)	0.8655	0.8896	0.8641	0.8774	0.8796	0.8891	0.9061	0.8085	0.8585	0.8491	0.8840	0.8713
Best value	0.8829	0.8948	0.8781	0.8897	0.8949	0.8970	0.9061	0.8374	0.8920	0.8601	0.8996	0.8850
Corpus-based IC models	Pearson correlation values for the IC-based similarity measures on the P&S _{full} dataset and WordNet 3.0											
ic-bnc-resnik-add1	0.8632	0.8821	0.8809	0.8853	0.8839	0.8291	0.8961	0.8230	0.8887	0.8629	0.8891	0.8734
ic-brown-resnik-add1	0.8673	0.8812	0.8749	0.8857	0.8783	0.8277	0.8942	0.8253	0.8888	0.8565	0.8873	0.8729
ic-semcor-add1	0.8692	0.8812	0.8746	0.8869	0.8779	0.8288	0.8927	0.8274	0.8900	0.8579	0.8875	0.8744
ic-semcorraw-add1	0.8792	0.8876	0.8825	0.8857	0.8855	0.8370	0.8983	0.8292	0.8899	0.8697	0.8968	0.8869
ic-semcorraw-resnik-add1	0.8772	0.8878	0.8805	0.8859	0.8839	0.8348	0.8991	0.8288	0.8897	0.8629	0.8946	0.8809
ic-shaks-resnik-add1	0.8620	0.8818	0.8784	0.8861	0.8809	0.8239	0.8961	0.8252	0.8879	0.8669	0.8966	0.8875
ic-treebank-add1	0.8736	0.8877	0.8749	0.8857	0.8778	0.8328	0.8988	0.8286	0.8909	0.8620	0.8922	0.8826
ic-treebank-resnik-add1	0.8704	0.8835	0.8736	0.8855	0.8770	0.8265	0.8973	0.8257	0.8893	0.8553	0.8901	0.8762
Best column value	0.8792	0.8878	0.8825	0.8869	0.8855	0.8370	0.8991	0.8292	0.8909	0.8697	0.8968	0.8875
Overall best value	0.8829	0.8948	0.8825	0.8897	0.8949	0.8970	0.9061	0.8374	0.8920	0.8697	0.8996	0.8875

the intrinsic Seco et al. IC model. Our results also prove that the classic Jiang–Conrath distance had not been properly exploited.

The correlation values reported by the hybrid IC-based measures do not justify their high computational cost. This conclusion applies to our weighted Jiang–Conrath measures (*coswj&C* and *wj&C*) and the similarity measures introduced in Li et al. (2003), Zhou et al. (2008b), Meng et al. (2014) and Gao et al. (2015). Our experimental results refute the practical use of any similarity measure based on features derived from the shortest path measures on a taxonomy. Therefore, we discard the use of these types of similarity measure in any practical application unless they are able to obtain correlation values much higher than less expensive

approaches, such as the *cosj&C* similarity measure, and the Meng et al. (2012) and Hadj Taieb et al. (2014a) measures.

In practice, there is no convincing winner from among the state-of-the-art similarity measures on WordNet and the problem is still open. Despite the new *coswj&C* similarity measure and the Hadj Taieb et al. (2014a) measure obtaining the best results in the RG65 and P&S_{full} datasets, all the evaluated measures, with the exception of the Meng et al. (2014) measure, obtain rivaling average correlation values of between 0.8599 and 0.8870. From this fact it follows that the performance for all the similarity measures evaluated herein is similar, thus, there is no convincing winner. This fact also endorses our previous conclusion as regards the hybrid IC-based measures.

Table 8
Correlation values for the IC-based similarity measures evaluated on the RG65 and Pirró datasets with WordNet 3.1.

Intrinsic IC models	State-of-the-art IC-based similarity measures									New measures (this work)		
	Pearson correlation values for the IC-based similarity measures on the RG65 dataset and WordNet 3.1											
	Resnik	Lin	J&C	Li_s9	Zhou	P&S	Meng12	Meng14	Gao	wj&C	cosj&C	coswj&C
Seco et al. (2004)	0.8326	0.8609	0.8546	0.8241	0.8708	0.8622	0.8596	0.8486	0.7988	0.8521	0.8642	0.8635
Zhou et al. (2008a, b)	0.8081	0.8260	0.8286	0.8438	0.8569	0.8334	0.8539	0.7749	0.7988	0.8273	0.8558	0.8566
Sánchez et al. (2011)	0.8409	0.8530	0.8619	0.8586	0.8638	0.8105	0.8663	0.8147	0.8682	0.8481	0.8751	0.8746
Sánchez et al. (2012)	0.8355	0.8616	0.8508	0.8615	0.8533	0.8332	0.8600	0.8475	0.8676	0.8505	0.8606	0.8615
Meng et al. (2012)	0.8260	0.8608	0.8598	0.8282	0.8631	0.8586	0.8670	0.8285	0.7988	0.8606	0.8723	0.8742
Yuan et al. (2013)	0.8243	0.8621	0.8505	0.8231	0.8616	0.8607	0.8675	0.8273	0.7988	0.8474	0.8632	0.8624
Best column value	0.8409	0.8621	0.8619	0.8615	0.8708	0.8622	0.8675	0.8486	0.8682	0.8606	0.8751	0.8746
Intrinsic IC models	Pearson correlation values for the IC-based similarity measures on the P&S _{full} dataset and WordNet 3.1											
Seco et al. (2004)	0.8799	0.8945	0.8781	0.8839	0.8940	0.8970	0.9032	0.8374	0.8583	0.8602	0.8966	0.8820
Zhou et al. (2008a, b)	0.8357	0.8420	0.8372	0.8897	0.8727	0.8563	0.8806	0.7488	0.8583	0.8231	0.8726	0.8585
Sánchez et al. (2011)	0.8740	0.8738	0.8762	0.8807	0.8788	0.8051	0.8964	0.7943	0.8844	0.8474	0.8996	0.8843
Sánchez et al. (2012)	0.8829	0.8948	0.8742	0.8877	0.8770	0.8398	0.9035	0.8351	0.8919	0.8588	0.8918	0.8791
Meng et al. (2012)	0.8646	0.8863	0.8715	0.8813	0.8804	0.8897	0.9025	0.8106	0.8583	0.8546	0.8917	0.8763
Yuan et al. (2013)	0.8655	0.8896	0.8641	0.8774	0.8792	0.8891	0.9061	0.8086	0.8583	0.8491	0.8840	0.8714
Best column value	0.8829	0.8948	0.8781	0.8897	0.8940	0.8970	0.9061	0.8374	0.8919	0.8602	0.8996	0.8843

5.1. Impact of the Wordnet version and datasets

We conclude that there are no significant differences in the correlation values obtained for the same measures in different WordNet versions. Looking at Table 4, we can see that in WordNet 3.0 and 3.1 the results are almost identical on both datasets, although between WordNet versions 2.1 and 3.x we can see a small difference of up to 0.060 for some measures.

Comparing the correlation values obtained by each measure on both datasets, we see a clear positive bias of 0.03 (3%) for the correlation values on the P&S_{full} dataset. We note that the P&S_{full} dataset shows a correlation value of 0.972 with regards to the RG65 dataset, thus, we can expect, as is shown, a similar difference between the correlation values obtained from both datasets. The bias for the P&S_{full} dataset is consistent for all the measures, although the percentage increase is not uniformly distributed in all the measures.

5.2. Intrinsic IC models versus corpus-based

Looking at Tables 6 and 7, we observe that the corpus-based IC models based on the Resnik method obtain rivaling results as regards the state-of-the-art intrinsic IC models. This conclusion refutes the accepted belief as to the clear superiority of the intrinsic IC models, as reported in Pirró (2009, Fig. 10) and Sánchez et al. (2012).

The results reported by Sánchez et al. are based on the benchmarks in Patwardhan and Pedersen (2006) and Pedersen (2010), which do not include the family of “add1” WordNet-based frequency files provided in Pedersen (2008). Despite this fact, we confirm that the state-of-the-art intrinsic IC models outperform the corpus-based ones, but with a much smaller margin. In practice, the corpus-based IC models evaluated herein obtain similar results, and we propose to select some of them as baseline for any future benchmark of IC-based similarity measures in WordNet.

These novel conclusions give rise to interesting questions as regards the relationship between the intrinsic and corpus-based IC models. From our point of view, we are now in a better position to evaluate the performance of the intrinsic IC models and understand the knowledge that they are producing. If an intrinsic IC model is able to mimic a corpus-based IC model, it means that in

some way, we would expect to find some relationship between the structure of the taxonomy, and the underlying corpus statistics.

5.3. Validation of previous methods

Our experiments confirm the results reported for the similarity measures introduced in Pirró and Seco (2008), Zhou et al. (2008b), Sánchez et al. (2012) and Hadj Taieb et al. (2014b,p. 256). Some of our implementation details and results were validated with the kind support of David Sánchez and Mohamed Hadj Taieb. The results reported in Li et al. (2003) cannot be compared directly because the parameter tuning and experimental conditions are different, however, the Li et al. measure obtains correlation values in the same range as the state of the art.

5.4. Contradictory results

On the other hand, we have found some contradictory results related to the results reported in Meng and Gu (2012), Meng et al. (2014), Hadj Taieb et al. (2014b) and Gao et al. (2015). In Meng and Gu (2012), the authors report a correlation value of 0.8804 in the RG65 dataset when their measure is combined with the intrinsic Seco et al. IC model. However, in Hadj Taieb et al. (2014b,p. 256) the authors report a value of 0.85 on WordNet 3.0, while herein, we report a correlation value of 0.8596. Therefore, we endorse the experimental results of Hadj Taieb et al. (2014b), in which the authors conclude that the Meng et al. (2012) measure performs better within the family of IC-based measures. In Meng et al. (2014), the same authors report a correlation value of 0.8817 with the Seco et al. IC model, however, in our experiments and herein, we obtain a correlation value of 0.8486. Indeed, the two results reported by Meng et al. are the highest correlation values reported in the literature for any intrinsic IC-based similarity measure in the RG65 dataset, however, we have not been able to confirm them. Finally, in a personal communication, the authors of Hadj Taieb et al. (2014b,p. 256) report a correlation value of 0.8784 for their similarity measure with adjustment factor ($sim_{taieb,2}$) on the RG65 dataset and WordNet 3.0, however, herein we obtain a value of 0.8597. These contradictory results confirm the difficulties in reproducing some previous approaches in the literature, a problem that is also noted in Fokkens et al. (2013). We invite the research

community to reproduce these previous approaches, as well as the results reported herein.

5.5. Warning about some reproducibility problems

Unlike the measures introduced in Meng and Gu (2012) and Meng et al. (2014), which are clearly described in their papers, the measure called strategy 3 in Gao et al. (2015) cannot be replicated exactly, because the authors omit the details on the IC model used in their experiments. In a series of personal communications, Jian-Bo Gao clarifies that they used the same corpus-based IC model as that used by Pedersen in Patwardhan and Pedersen (2006), however, he does not say which IC file was used, and he did not provide us with their IC model files to be able replicate their experiments exactly. Herein, we have evaluated the strategy 3 measure in Gao et al. (2015) with a large set of corpus-based IC models built from the dataset published in Pedersen (2008), and we have not obtained the same results.

We think that these contradictory results could be derived from some minor implementation differences, and it represents a good opportunity to work on the reproducibility problems in the area. As you can see in Table 4, the performance margin for all the methods is very narrow, due to the huge complexity of the underlying cognitive problem, and that the state of the problem has given rise to an asymptotic behavior within the range (0.87, 0.88).

The reproducibility of intrinsic IC models and similarity measures is a difficult problem, which has been eluded in many works where the authors cite and take valid results obtained by others. The implementation of similarity measures requires sensitive graph-based algorithms to compute all types of taxonomical features, which are not unambiguously defined in taxonomies with a multiple inheritance such as WordNet. For example, the depth of a concept could be defined as the longest or shortest ascending path from any node to the root node. With the exception of the work in Hadj Taieb et al. (2014b, Section 4.1), most authors do not clarify this issue. The same argument is valid for the definition of the lowest concept subsumer (LCS), also called lowest common ancestor (LCA), which is only well defined for IC-based similarity measures through the introduction of the most informative common ancestor (MICA) notion.

In Fokkens et al. (2013), the authors also warn about the reproduction problems in the evaluation of the semantic similarity measures, and the need to validate previous methods and experiments, claims which we also make in this work. Most recent works in the area cite results in previous works, without replicating them. The replication of previous methods is complex, it requires a considerable effort for the implementation and recovery of missing details, and it is not exempt from risk. However, the reproducibility of the published results is an essential feature of science, and we can learn so much from this process. Therefore, we invite the research community to replicate the methods and experiments introduced in this work, with the aim of validating previously reported results.

6. Conclusions and future work

First, we have introduced one IC-based semantic distance and three new IC-based similarity measures based on a generalization and normalization of the classic Jiang–Conrath distance, which outperform the state-of-the-art methods in the RG65 dataset. Second, we introduce an up-to-date experimental survey, whose aim is the uniform comparison of the most recent and relevant similarity measures on WordNet, especially the families of IC-based similarity measures and intrinsic IC models. In addition, we

introduce an experimental comparison between the intrinsic and corpus-based IC model that allows some previous conclusions to be refuted.

We confirm that the state-of-the-art on similarity measures is led by the family of IC-based measures, specifically by our new cosine-normalized measures and the Meng et al. (2012) similarity measure. This latter statement is also endorsed by the best Hadj Taieb et al. measure, because despite it not being based on an IC model, this measure is inspired by and closely related to the Seco et al. IC model. Our experimental results allow the use of any hybrid IC-based measure to be refuted due to their high computational cost for a similar performance.

Despite of the corpus-based IC models evaluated herein obtaining rivaling results as regards the state-of-the-art intrinsic IC models, we confirm that the intrinsic IC models slightly outperform the former ones. However, the difference between the corpus-based IC models and the intrinsic ones is smaller than that reported in the literature, which was based on corpus-based IC models built with the Resnik method on other WordNet-based frequency files.

Finally, we confirm that there is no significant difference in the performance of the similarity measures in different WordNet versions. We also subscribe to the warning made in Fokkens et al. (2013) on the reproducibility problems related to the similarity measures in WordNet, and we also invite to the research community to replicate previous approaches and experiments in their future research.

As forthcoming activities, we would like to study the relationship between the corpus-based IC models evaluated herein and the intrinsic IC models. We would also like to study the integration of ontology-based similarity measures with concept-based distributional methods. Another interesting line of research is the proposal and promotion of an open framework for the exact reproducibility of the similarity measures and benchmarks reported in the literature.

Acknowledgements

Despite deciding to develop our own software library to implement all the IC-based models and measures evaluated in this work, we would like to express our gratitude to Sébastien Harispe, who even provided us the source code of the SML library, offering his total support. Jian-Bo Gao, David Sánchez, Montserrat Batet and Giuseppe Pirró kindly answered all our questions to clarify certain issues on their methods and experimental results to replicate them in our platform. Mohamed Hadj Taieb kindly offered us his total support in replicating their similarity measures exactly. Ted Pedersen kindly answered all our questions and provided us with the WordNet-based frequency files used to build all the corpus-based IC models used in our experiments. Alexis Moreno-Pulido helped us to find some old papers. Mark Hallett reviewed the english translation. Finally, we are very grateful for the comments made by the anonymous reviewers to improve the quality of the paper. To all of them, we would like to show our most sincere gratitude. This work has been partially supported by the Spanish VOXPOPULI Project (TIN2013-47090-C3-1-P).

Appendix A. Result tables for the IC-based similarity measures

See Tables 5,6,7,8.

References

- Alvarez, M.A., Lim, S., 2007. A graph modeling of semantic similarity between words. In: Proceedings of the First IEEE International Conference on Semantic Computing (ICSC 2007). IEEE Computer Society, Irvine, California USA, September, pp. 355–362.
- Batet, M., Sánchez, D., Valls, A., 2011. An ontology-based measure to compute semantic similarity in biomedicine. *J. Biomed. Inform.* 44 (February (1)), 118–125.
- Budanitsky, A., Hirst, G., 2006. Evaluating WordNet-based measures of lexical semantic relatedness. *Comput. Linguist.* 32 (March (1)), 13–47.
- Chan, L.W.-C., Liu, Y., Shyu, C.-R., Benzie, I.F.F., 2011. A SNOMED supported ontological vector model for subclinical disorder detection using EHR similarity. *Eng. Appl. Artif. Intell.* 24, 1398–1409.
- Chen, M., Chowdhury, R.A., Ramachandran, V., Roche, D.L., Tong, L., 2007. Priority queues and Dijkstra's algorithm. Technical Report TR-07-54, Computer Science Department, University of Texas at Austin.
- Couto, F.M., Pinto, H.S., 2013. The next generation of similarity measures that fully explore the semantics in biomedical ontologies. *J. Bioinform. Comput. Biol.* 11 (October (5)), 1371001.
- Cross, V., Hu, X., 2011. Using semantic similarity in ontology alignment. In: Proceedings of the Sixth International Workshop on Ontology Matching (OM), 10th International Semantic Web Conference (ISWC 2011), Bonn Germany pp. 61–72.
- Cross, V., Yu, X., Hu, X., 2013. Unifying ontological similarity measures: a theoretical and empirical investigation. *Int. J. Approx. Reason.* 54 (September (7)), 861–875.
- Fiorini, N., Ranwez, S., Montmain, J., Ranwez, V., 2015. USI: a fast and accurate approach for conceptual document annotation. *BMC Bioinform.* (16:83), 14 March.
- Fokkens, A., Van Erp, M., Postma, M., Pedersen, T., Vossen, P., Freire, N., 2013. Offspring from reproduction problems: what replication failure teaches us. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. ACL, Sofia Bulgaria, 4 August, pp. 1691–1701.
- Gabrilovich, E., Markovitch, S., 2007. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In: Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI-07), vol. 7. Morgan Kaufmann Publishers Inc., Hyderabad, India, pp. 1606–1611.
- Gan, M., Dou, X., Jiang, R., 2013. From ontology to semantic similarity: calculation of ontology-based semantic similarity. *Sci. World J.* 2013, 793091.
- Gao, J.-B., Zhang, B.-W., Chen, X.-H., 2015. A WordNet-based semantic similarity measurement combining edge-counting and information content theory. *Eng. Appl. Artif. Intell.* 39 (March (0)), 80–88.
- Garla, V.N., Brandt, C., 2012. Semantic similarity in the biomedical domain: an evaluation across knowledge sources. *BMC Bioinform.* 13, 261, October.
- Hadj Taieb, M.A., ben Aouicha, M., ben Hamadou, A., 2014a. A new semantic relatedness measurement using WordNet features. *Knowl. Inf. Syst.* 41 (November (2)), 467–497.
- Hadj Taieb, M.A., ben Aouicha, M., ben Hamadou, A., 2014b. Ontology-based approach for measuring semantic similarity. *Eng. Appl. Artif. Intell.* 36 (November (0)), 238–261.
- Harispe, S., Ranwez, S., Janaqi, S., Montmain, J., 2014a. The semantic measures library: assessing semantic similarity from knowledge representation analysis. In: Métais, E., Roche, M., Teisseire, M. (Eds.), *Natural Language Processing and Information Systems. Proceedings of the 19th International Conference on Applications of Natural Language to Information Systems*. vol. 8455 of LNCS. Springer, Montpellier, France, 18 June, pp. 254–257.
- Harispe, S., Ranwez, S., Janaqi, S., Montmain, J., 2015. Semantic similarity from natural language and ontology analysis. vol. 8 of *Synthesis Lectures on Human Language Technologies*. Morgan & Claypool publishing, May.
- Harispe, S., Sánchez, D., Ranwez, S., Janaqi, S., Montmain, J., 2014. A framework for unifying ontology-based semantic similarity measures: a study in the biomedical domain. *J. Biomed. Inform.* 48, 38–53, April.
- Harris, Z.S., 1981. Distributional structure. In: Hiz, H. (Ed.), *Papers on Syntax*. vol. 14 of *Synthese Language Library*. Springer, Netherlands, pp. 3–22.
- Hirst, G., St-Onge, D., 1998. Lexical chains as representations of context for the detection and correction of malapropisms. In: Fellbaum, C. (Ed.), *WordNet: An electronic lexical database*. Massachusetts Institute of Technology, pp. 305–332.
- Jiang, J.J., Conrath, D.W., 1997. Semantic similarity based on corpus statistics and lexical taxonomy. In: Proceedings of International Conference Research on Computational Linguistics (ROCLING X). pp. 19–33.
- Lastra-Díaz, J.J., 2014. Intrinsic semantic spaces for the representation of documents and semantic annotated data. Master's thesis, Universidad Nacional de Educación a Distancia (UNED). Department of Computer Languages and Systems, 29 September. <<http://e-spacio.uned.es/fez/view/bibliuned:mater-ETSIIinformatica-LSI-Jlastra>>.
- Lastra-Díaz, J.J., García-Serrano, A., 2014. System and method for the indexing and retrieval of semantically annotated data using an ontology-based information retrieval model. United States Patent and Trademark Office (USPTO) application US14/576,679, 19 December.
- Leacock, C., Chodorow, M., 1998. Combining local context and WordNet similarity for word sense identification. In: Fellbaum, C. (Ed.), *WordNet: an electronic lexical database*. Massachusetts Institute of Technology, pp. 265–283.
- Li, P., Wang, H., Zhu, K.Q., Wang, Z., Hu, X.-G., Wu, X., 2015. A large probabilistic semantic network based approach to compute term similarity. *IEEE Trans. Knowl. Data Eng.* 0 1–14, in press.
- Li, P., Wang, H., Zhu, K.Q., Wang, Z., Wu, X., 2013. Computing term similarity by large probabilistic is a knowledge. In: Proceedings of the 22nd ACM International Conference on Conference on Information & Knowledge Management. (CIKM '13). ACM, New York, NY, USA, pp. 1401–1410.
- Li, Y., Bandar, Z.A., McLean, D., 2003. An approach for measuring semantic similarity between words using multiple information sources. *IEEE Trans. Knowl. Data Eng.* 15 (4), 871–882.
- Lin, D., 1998. An information-theoretic definition of similarity. In: Proceedings of the 15th International Conference on Machine Learning, vol. 98, Madison, WI, pp. 296–304.
- Meng, L., Gu, J., 2012. A new model for measuring word sense similarity in WordNet. In: Proceedings of the 4th International Conference on Advanced Communication and Networking, ASTL, vol. 14, pp. 18–23.
- Meng, L., Gu, J., Zhou, Z., 2012. A new model of information content based on concept's topology for measuring semantic similarity in WordNet. *Int. J. Grid Distrib. Comput.* 5 (September (3)), 81–93.
- Meng, L., Huang, R., Gu, J., 2014. Measuring semantic similarity of word pairs using path and information content. *Int. J. Futur. Gener. Commun. & Netw.* 7 (June (3)), 183–194.
- Mihalcea, R., Corley, C., Strapparava, C., 2006. Corpus-based and knowledge-based measures of text semantic similarity. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 1, AAAI Press, pp. 775–780.
- Miller, G.A., 1995. WordNet: a lexical database for English. *Commun. ACM* 38 (11), 39–41.
- Mohammad, S., Hirst, G., 2006. Distributional measures of concept-distance: a task-oriented evaluation. In: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing. EMNLP '06. Association for Computational Linguistics, Stroudsburg, PA USA, pp. 35–43.
- Mohammad, S.M., Hirst, G., 2012. Distributional measures of semantic distance: a survey, 8 March <http://arxiv:1203.1858> arxiv:1203.1858.
- Orum, C., Joslyn, C.A., 2009. Valuations and metrics on partially ordered sets, March <http://arxiv:0903.2679> arxiv:0903.2679.
- Patwardhan, S., Banerjee, S., Pedersen, T., 2003. Using measures of semantic relatedness for word sense disambiguation. In: Gelbukh, A. (Ed.), Proceedings of the 4th International Conference on Computational Linguistics and Intelligent Text Processing (CICLING 2003), vol. 2588 of LNCS, Springer, Mexico D.F., February, pp. 241–257.
- Patwardhan, S., Pedersen, T., 2006. Using WordNet-based context vectors to estimate the semantic relatedness of concepts. In: Proceedings of the EACL 2006 Workshop Making Sense of Sense-Bringing Computational Linguistics and Psycholinguistics Together, vol. 1501, Trento, Italy, pp. 1–8.
- Pedersen, T., 2008. WordNet-InfoContent-3.0.tar dataset repository. https://www.researchgate.net/publication/273885902_WordNet-InfoContent-3.0.tar.
- Pedersen, T., 2010. Information content measures of semantic similarity perform better without sense-tagged text. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. HLT '10. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 329–332.
- Pedersen, T., 2013. Measuring the similarity and relatedness of concepts: a MICAI 2013 tutorial. Tutorial presentation within the 12th Mexican international conference on artificial intelligence, 25 November. <http://dx.doi.org/10.13140/RG.2.1.3025.6164>.
- Pesquita, C., Faria, D., Falcao, A.O., Lord, P., Couto, F.M., 2009. Semantic similarity in biomedical ontologies. *PLoS Comput. Biol.* 5 (7), e1000443.
- Pirró, G., 2009. A semantic similarity metric combining features and intrinsic information content. *Data Knowl. Eng.* 68 (November (11)), 1289–1308.
- Pirró, G., Euzenat, J., 2010. A feature and information theoretic framework for semantic similarity and relatedness. In: Patel-Schneider, P.F., Pan, Y., Hitzler, P., Mika, P., Zhang, L., Pan, J.Z., Horrocks, I., Glimm, B. (Eds.), In: Proceedings of the 9th International Semantic Web Conference, ISWC 2010, vol. 6496 of LNCS, Springer, Shanghai, China, 7 November, pp. 615–630.
- Pirró, G., Seco, N., 2008. Design, implementation and evaluation of a new semantic similarity metric combining features and intrinsic information content. In: Meersman, R., Tari, Z. (Eds.), *On the Move to Meaningful Internet Systems: OTM 2008*. vol. 5332 of LNCS. Springer, January, pp. 1271–1288.
- Rada, R., Mili, H., Bicknell, E., Blettner, M., 1989. Development and application of a metric on semantic nets. *IEEE Trans. Syst. Man Cybern.* 19 (1), 17–30.
- Resnik, P., 1995. Using information content to evaluate semantic similarity in a taxonomy. In: Proceedings of the International Joint Conferences on Artificial Intelligence (IJCAI 1995). vol. 1, Montreal, Canada, 20 August, pp. 448–453.
- Resnik, P., 1999. Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language. *J. Artif. Intell. Res.* 11, 95–130, July.
- Rubenstein, H., Goodenough, J.B., 1965. Contextual correlates of synonymy. *Commun. ACM* 8 (October (10)), 627–633.
- Sánchez, D., Batet, M., 2012. A new model to compute the information content of concepts from taxonomic knowledge. *Int. J. Semant. Web Inf. Syst.* 8, 2.
- Sánchez, D., Batet, M., Isern, D., 2011. Ontology-based information content computation. *Knowl.-Based Syst.* 24 (March (2)), 297–303.
- Sánchez, D., Batet, M., Isern, D., Valls, A., 2012. Ontology-based semantic similarity: a new feature-based approach. *Expert Syst. Appl.* 39, 7718–7728.

- Sánchez, D., Batet, M., Martínez, S., Domingo-Ferrer, J., 2015. Semantic variance: an intuitive measure for ontology accuracy evaluation. *Eng. Appl. Artif. Intell.* 39 (March (0)), 89–99.
- Saruladha, K., Aghila, G., Raj, S., 2010. A survey of semantic similarity methods for ontology based information retrieval. In: *Proceedings of the Second International Conference on Machine Learning and Computing (ICMLC 2010)*, IEEE, pp. 297–301.
- Sebt, A., Barfroush, A.A., 2008. A new word sense similarity measure in WordNet. In: *Proceedings of the International Multiconference on Computer Science and Information Technology, IMCSIT 2008*, IEEE, October, pp. 369–373.
- Seco, N., Veale, T., Hayes, J., 2004. An intrinsic information content metric for semantic similarity in WordNet. In: López de Mántaras, R., Saitta, L. (Eds.), *Proceedings of the 16th European Conference on Artificial Intelligence (ECAI)*, vol. 16. IOS Press, Valencia, Spain, pp. 1089–1094.
- Turney, P.D., Pantel, P., 2010. From frequency to meaning: vector space models of semantics. *J. Artif. Intell. Res.* 37, 141–188.
- Tversky, A., 1977. Features of similarity. *Psychol. Rev.* 84 (July (4)), 327–352.
- Wu, Z., Palmer, M., 1994. Verbs semantics and lexical selection. In: *Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics. ACL '94*. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 133–138.
- Yan, W., Zanni-Merk, C., Cavallucci, D., Collet, P., 2014. An ontology-based approach for inventive problem solving. *Eng. Appl. Artif. Intell.* 27 (January (0)), 175–190.
- Yuan, Q., Yu, Z., Wang, K., 2013. A new model of information content for measuring the semantic similarity between concepts. In: *Proceedings of the International Conference on Cloud Computing and Big Data (CloudCom-Asia 2013)*. IEEE Computer Society, December, pp. 141–146.
- Zhou, Z., Wang, Y., Gu, J., 2008a. A new model of information content for semantic similarity in WordNets. In: *Proceedings of the Second International Conference on Future Generation Communication and Networking Symposia (FGCNS'08)*, vol. 3. IEEE, pp. 85–89.
- Zhou, Z., Wang, Y., Gu, J., 2008b. New model of semantic similarity measuring in WordNet. In: *Proceedings of the 3rd International Conference on Intelligent System and Knowledge Engineering (ISKE 2008)*, vol. 1. IEEE, November, pp. 256–261.

Chapter 8

Knowledge-Based systems article

This page intentionally left blank.



A new family of information content models with an experimental survey on WordNet



Juan J. Lastra-Díaz*, Ana García-Serrano

NLP & IR Research Group, E.T.S.I. Informática – UNED, Universidad Nacional de Educación a Distancia, C/Juan del Rosal 16, 28040 Madrid, Spain

ARTICLE INFO

Article history:

Received 27 February 2015
 Received in revised form 25 August 2015
 Accepted 30 August 2015
 Available online 5 September 2015

Keywords:

Semantic similarity
 Intrinsic information content model
 Ontology-based semantic similarity measures and distances
 IC-based similarity measures
 Semantic similarity benchmark based on WordNet

ABSTRACT

This paper introduces a new family of intrinsic and corpus-based Information Content (IC) models for ontology-based similarity measures based on the IC theory, a detailed state of the art, an experimental survey of IC models and IC-based similarity measures on WordNet, and a comparison between intrinsic and corpus-based IC models. The family of IC models is made up of five intrinsic IC models, called *CondProbHypo*, *CondProbUniform*, *CondProbLeaves*, *CondProbLogistic*, and *CondProbCosine*, and one corpus-based IC model called *CondProbCorpus* which completes the family. The proposed IC models rely on two previously unconsidered notions: (1) the preservation of the probabilistic structure of the taxonomy associated to the conditional probabilities between child and parent concepts, and (2) the explicit consideration of a cognitive similarity notion in the definition of the IC model. The family of IC models defines a new method for the proposal of new intrinsic IC models based on the exploration of other alternatives for the intrinsic estimation of the conditional probabilities between child and parent concepts. Our work is inspired by an unexplored relationship between the Jiang–Conrath distance and a shortest path on an IC-based weighted graph, derived from the conditional probabilities between concepts, as well as certain cognitive evidence about the perception distance between concepts. The new IC models obtain results comparable to the state of the art and satisfy a set of well-founded structure axioms. In addition, we prove that most of intrinsic IC models and IC-based similarity measures do not show a significant statistical difference as regards a baseline corpus-based IC model and the Jiang–Conrath similarity, with the exception of the overall outperformance shown by the Sánchez et al. (2012) IC model and the *cos*/&C similarity measure, which has recently been introduced by the authors.

© 2015 Elsevier B.V. All rights reserved.

1. Introduction

The ontology-based similarity measures have found many applications in the fields of natural language processing (NLP), information retrieval (IR), and bioengineering. Many semantic tasks require the evaluation of the degree of similarity between words and concepts, as perceived by a human being. For instance, in Lastra-Díaz [24], we introduce an ontology-based IR model, called *Intrinsic Ontological Spaces*, which is based on a metric space defined by a generalization of the Jiang–Conrath distance to populated ontologies, in which an ontology-based semantic distance, called the *weighted Jiang–Conrath distance*, is used to define a metric space of weighted mentions to classes and individuals. In Mihalcea et al. [34], the authors propose a text similarity measure based on the combination of an IDF weighting scheme with any

ontology-based similarity measure. In Patwardhan et al. [37], the authors introduce a word sense disambiguation (WSD) method based on the distributional hypothesis and the use of ontology-based similarity measures to select the closest evocated concept between a disambiguated word and its neighboring words. The ontology-based similarity measures have also been applied to the ontology alignment (OA) problem Zong et al. [65]. For instance, in Cross and Hu [8], the authors review the use of semantic similarity measures on the ontology alignment (OA) problem and introduce a semantic alignment quality measure based on the difference between the similarity functions defined at the input and on target ontologies. In another work Wang et al. [59], the authors introduce a tree mapping algorithm defined as an optimal search problem based on a binary tree-valued similarity measure derived from three ontology-based similarity measures. In the theory of inventive problems Yan et al. [61], the authors propose a method to solve the inventive problem based on the definition of a short text similarity measure as the maximum of the pairwise Lin similarity measure between concept sets. In the field of

* Corresponding author.

E-mail addresses: jlastra@invi.uned.es (J.J. Lastra-Díaz), agarcia@lsi.uned.es (A. García-Serrano).

bioengineering Sánchez and Batet [50], the authors propose a reformulation of some known ontology-based similarity measures, which relies on the length of shortest path between concepts in terms of an IC model. In Couto and Pinto [7] and Pesquita et al. [42], the authors survey other applications of ontology-based semantic similarity measures in bioengineering, such as the prediction of protein functions.

From an abstract standpoint, any comparison of concepts requires the evocation of a common conceptual model where the comparison takes place. Much cognitive evidence suggests that human beings organize knowledge in a hierarchical manner, through a categorization process of the reality. It is a known fact that the similarity notion is mainly encoded by the “is-a” relationships within a taxonomy. Any taxonomy on a set of concepts C is defined formally by a triplet $\mathcal{C} = (C, \leq_C, \Gamma)$, where (C, \leq_C) is a partially ordered set, and $\Gamma \in C$ is a distinguished supreme element called the root, such that $\forall c_i \in C \rightarrow c_i \leq_C \Gamma$. In this way, the ontology-based semantic similarity problem can be formulated as follows: given a taxonomy of concepts $\mathcal{C} = (C, \leq_C, \Gamma)$ and two input words evocating two sets of concepts in C , find a binary function $sim : C \times C \rightarrow \mathbb{R}$, that approximates, as well as possible the degree of similarity as perceived by a human being. In this work, we focus in the study of ontology-based similarity measures and IC models defined on a single taxonomy, which is based solely on “is-a” relationships, despite other works, such as Li et al. [29] and Pirró and Euzenat [44], also consider “part-of” relationships.

In some applications, the availability of a single ontology to compute similarity measures could be a serious drawback, especially for applications requiring vocabularies and concepts from different technical domains. This latter problem has given rise to the proposal of methods for the estimation of semantic similarity measures combining multiple ontologies, such as the method for feature-based measures proposed in Solé-Ribalta et al. [57], as well as the method for IC-based similarity measures proposed in Batet et al. [4]. This latter work is especially relevant to our work due

to its direct application to the IC models and similarity measures studied herein.

The taxonomy-based similarity estimation is a very old problem, which has been researched since the nineteen-seventies in different fields, ranging from cognitive psychology Tversky [58], to information retrieval Rada et al. [46]. Most of works categorize the different ontology-based semantic measures into three families, although there are also hybrid approaches, as follows: (1) edge-counting measures, whose pioneering work has been carried out by Rada et al. [46], (2) IC-based measures, whose main references are Resnik [47], Jiang and Conrath [23] and Lin [30], and the (3) feature-based measures, whose pioneering work has been carried out by Tversky [58], and the most recent one is Sánchez et al. [53].

The state of the art in ontology-based semantic similarities and distances is defined by the family of intrinsic IC-based measures, which are defined by the combination of one intrinsic IC model with any IC-based measure. This statement is endorsed by several WordNet-based benchmarks in the literature, such as Budanitsky and Hirst [6], Sánchez et al. [52], Pirró [43], Hadj Taieb et al. [15] and Lastra-Díaz and García-Serrano [26]. Despite there being some relevant non IC-based similarity measures in the literature, such as the feature-based measure proposed in Sánchez et al. [53], and the hyponym-based approach proposed in Hadj Taieb et al. [15], the mainstream is still the proposal of new intrinsic IC-based models and measures, such as the works in Pirró and Euzenat [44], Meng et al. [33], Gao et al. [12], and Lastra-Díaz and García-Serrano [26].

The IC-based similarity measures need a concept-valued function, called the IC model, which defines the IC value for each concept within the ontology. Given a taxonomy of concepts $\mathcal{C} = (C, \leq_C, \Gamma)$, an information content model is a concept-valued function $IC : C \rightarrow \mathbb{R}^+ \cup \{0\}$, which represents an estimation of the information content for every concept $c_i \in C$, defined by $IC(c_i) = -\log_2(p(c_i))$, where $p(c_i)$ is the occurrence probability of the concept c_i . Once the IC-based measure is chosen, the IC model

Table 1
State-of-the-art ontology-based similarity measures evaluated in our experiments.

Reference	Definition of the IC-based similarity measures
Resnik [47]	$sim_{Resnik}(c_1, c_2) = IC(MICA(c_1, c_2))$
Jiang and Conrath [23]	$d_{J&C}(c_1, c_2) = IC(c_1) + IC(c_2) - 2IC(MICA(c_1, c_2))$ $sim_{J&C}(c_1, c_2) = 1 - \frac{1}{2}d_{J&C}(c_1, c_2)$
Lin [30]	$sim_{Lin}(c_1, c_2) = \frac{2IC(MICA(c_1, c_2))}{IC(c_1) + IC(c_2)}$
Pirró and Seco [45]	$sim_{P\&S}(c_1, c_2) = \begin{cases} \frac{3IC(MICA(c_1, c_2))}{-IC(c_1) - IC(c_2)}, & \text{if } c_1 \neq c_2 \\ 1, & \text{if } c_1 = c_2 \end{cases}$
Pirró and Euzenat [44]	$sim_{FaTH}(c_1, c_2) = \frac{IC(MICA(c_1, c_2))}{IC(c_1) + IC(c_2) - IC(MICA(c_1, c_2))}$
Meng and Gu [31]	$sim_{Meng}(c_1, c_2) = e^{sim_{Lin}(c_1, c_2)} - 1 = e^{\frac{2IC(MICA(c_1, c_2))}{IC(c_1) + IC(c_2)}} - 1$
Lastra-Díaz and García-Serrano [26]	$sim_{cosJ\&C}(c_1, c_2) = 1 - \cos\left(\frac{\pi}{2} \left(1 - \frac{d_{J\&C}(c_1, c_2)}{2 \times \max_{J\&C}}\right)\right)$ $\max_{d_{J\&C}} = \max_{c \in Leaves(C)} \{IC(c)\}$
Reference	Definition of the non IC-based similarity measures
Hadj Taieb et al. [15]	$sim_{Taieb}(c_1, c_2) = \frac{ TermDepth(c_1, c_2) \times TermHypo(c_1, c_2)}{2 \times \frac{depth(c_1, c_2)}{depth(c_1) + depth(c_2)} \times \frac{SpecHypo(c_1, c_2)}{SpecHypo(c_1, c_2) + SpecHypo(c_1, c_2)}}$ $TermDepth(c_1, c_2) = \frac{2 \times depth(c_1, c_2)}{depth(c_1) + depth(c_2)}$ $TermHypo(c_1, c_2) = \frac{2 \times SpecHypo(c_1, c_2)}{SpecHypo(c_1, c_2) + SpecHypo(c_1, c_2)}$ $SpecHypo(c_1, c_2) = 1 - \frac{\log(HypoValue(c))}{\log(HypoValue(root))}$ $HypoValue(c) = \sum_{c' \in HypoInc(c)} P(depth(c'))$ $P(depth(c')) = \frac{ \{c' \in C depth(c') = depth(c)\} }{ C }$ $depth(c) = \text{length of the longest ascending path to root}$ $HypoInc(c) = \{c' \in C c' \leq c\}$

is the mainly responsible for the definition of the notion of similarity and distance between concepts. Therefore, each IC model defines the underlying semantic metric of the base taxonomy, when it is interpreted as a metric space. Another application of the IC models is the definition of semantic relatedness measures. For instance, in Pirró and Euzenat [44], the authors introduce an extended IC model which integrates “part-of” relationships. Then, the extended IC model is combined with the sim_{FATH} similarity measure shown in Table 1 to define a relatedness measure.

The IC models are categorized into two main groups according to the information source used in their computation: (a) corpus-based IC models, such as that proposed in Resnik [48], and (b) intrinsic IC models, which are based on the information encoded into the ontology structure. The main drawback of the corpus-based IC models is the difficulty of getting a well-balanced and disambiguated corpus for most conceptual models. This latter fact has encouraged the development of intrinsic IC models, such as the pioneering work in [56], whose core hypothesis is that the IC values can be estimated directly from the structure of the taxonomy. Thus, the development of new intrinsic IC-based measures is divided into two closely related problems: (1) the proposal of new intrinsic IC models, as in our work, and (2) the proposal of new IC-based similarity measures.

1.1. Main motivation and hypothesis

The main motivation for this work is our observation that the conditional probability functions encode some structure axioms that should be verified by any IC model, but the IC models in the literature do not consider them, with the exception of the work of Sebtí and Barfroush [55]. Our main hypothesis is that the explicit encoding of these structure axioms in the IC models should lead us to improve the performance of the IC models in semantic similarity tasks, and to a better understanding of the problem.

A second motivation for our work is a first try at integrating some ideas in cognitive psychology into the IC models. In [13, section 2.8], Gärdenfors introduces a conceptual space model based on a Voronoi diagram, with the aim of explaining a number of plausible production mechanisms for the vagueness of concepts and their categorical perception. However, Gärdenfors does not provide a specific metric for this space, which is precisely the aim of this paper, and the whole family of intrinsic IC models. In [13, p.46], the author notes that the mechanisms that explain the vagueness notion, also explain another phenomenon in the cognitive perception of categories which can be defined as follows: *the instance of a concept is more quickly perceived as belonging to another category, when the distance from the prototype of the category increases*. This last idea can be formulated through the definition of the *cognitive similarity function as a non-linear function of sigmoid type over the underlying metric of the conceptual space*, which led us to the core idea behind the cognitive IC models called *CondProbLogistic* and *CondProbCosine*.

1.2. Definition of the problem and contributions

The main aim of this paper is to introduce a new family of intrinsic and corpus-based IC models that share a common computational and algebraic structure. The family includes five intrinsic IC models, called *CondProbHypo*, *CondProbUniform*, *CondProbLeaves*, *CondProbLogistic* and *CondProbCosine*, and one corpus-based IC model that completes the family, called *CondProbCorpus*. The proposed IC models are based on two previously unconsidered notions: (1) the preservation of the probabilistic structure of the taxonomy, encoded by the edge-based conditional probability, and (2) the explicit modeling of a cognitive similarity notion within

the IC model. Our work belongs to the family of intrinsic IC models for IC-based semantic measures, and it is inspired by an unexplored relationship between the Jiang–Conrath distance and the length of the shortest path between concepts on an IC-based weighted graph, as well as some remarks from Gärdenfors on the perception distance between concepts.

In addition, the work includes other significant contributions. First, we carried-out a large benchmark of IC-based models and similarity measures on WordNet 3.0, which is based on our own code implementation in order to replicate previous methods and results reported in the literature. The experiments include the five most relevant datasets, thirteen intrinsic IC models, four corpus-based IC models, the eight best IC-based similarity measures reported in Lastra-Díaz and García-Serrano [26], and one intrinsic non IC-based measure introduced in Hady Taieb et al. [15]. Third, we introduce a new comparison between intrinsic and corpus-based IC models that allows some previous conclusions on the out-performance of the intrinsic IC models over the corpus-based to be refuted. This latter finding confirms a similar finding in our aforementioned work in a more conclusive manner, in which we prove that “the margin of performance between the intrinsic and corpus-based IC models is much smaller than the research community first thought”. Fourth, encouraged by the latter finding we propose a new baseline for IC models defined by two corpus-based IC models based on an unexplored WordNet-based frequency file, and the Resnik and *CondProbCorpus* IC models.

The rest of the paper is structured as follows. In Section 2, we review the literature on ontology-based similarity measures, especially those evaluated in this study. Section 3 is devoted to reviewing the literature on intrinsic IC models. Section 4 introduces the new family of IC models. In Section 5, we describe the evaluation methodology and the results obtained. Section 6 introduces our discussion of the results. Finally, we present our conclusions and future work.

2. Ontology-based similarity and distance measures

The literature on ontology-based semantic similarity measures and distances is very extensive. Herein, we focus in the family of IC-based measures and models, in which our work is framed. For a broader and recent survey, we refer the reader to the book of Harispe et al. [18], and our state of the art in Lastra-Díaz and García-Serrano [26]. Other general surveys can be found in Saruladha et al. [54] and Sánchez et al. [53], whilst in the field of bioengineering, we find the works in Pesquita et al. [42], Hsieh et al. [22], Cross et al. [9], and Harispe et al. [19].

Modern research into the area starts with the work in Rada et al. [46]. In this work, the authors propose to use the length of the shortest path between concepts on an ontology as a distance measurement between them. Their core hypothesis is that the conceptual distance, or similarity, between concepts in a semantic network, is proportional to the length of the path that links them. Other subsequent works propose different similarity measures based on the integration of the length of the shortest path, such as Lee et al. [28], Wu and Palmer [60], Leacock and Chodorow [27] and Hirst and St-Onge [21].

The main drawback of the measures in the edge-counting family above, called the *uniform weighting* premise herein, is that they implicitly assume that every edge has the same relevance in the computation of the overall length of the path, without considering the depth or occurrence probability of the concepts. In Resnik [47], the author proposes a new semantic similarity based on an Information Content (IC) measure, whose main motivation is to remove the *uniform weighting* premise of the edge-counting measures. The basic hypothesis behind all the IC-based similarity measures is that the more abstract concepts should have lower information content

than the more specific ones, and the higher the conditional probability between any concept and its parent, the lower its distance. According to Resnik, the IC measure for every concept $c_i \in C$ is the negative logarithm of its occurrence probability $p(c_i)$, as defined in Eq. (1). Resnik defines the similarity measure between two concepts as the IC value of the most informative common ancestor (MICA), as shown in Table 1.

$$IC(c_i) = -\log_2(p(c_i)) \quad (1)$$

One drawback of the Resnik similarity measure is that it only considers the IC value of the lowest common ancestor concept, not the information along the path between concepts. With the aim of bridging this gap, in Jiang and Conrath [23] the authors introduce the IC-based semantic distance denoted by $d_{J\&C}$ in Table 1, whose similarity version, denoted $sim_{J\&C}$, is defined by a negative linear transformation. With the same aim, in Lin [30] the author introduces the similarity measure denoted by sim_{lin} in Table 1, which could be interpreted as an IC-based formulation of the set-based Dice coefficient.

In Tversky [58], the authors introduce the first feature-based semantic similarity measure, which is defined by a weighted variant for the complement of the symmetric difference between the feature set of two concepts. Despite the meaning of the Tversky measure is clear and well-founded, this measure can only be used whether the feature sets for all the concepts within a taxonomy are known, which is a strong limitation for its practical use. With the aim to bridge the gap in the Tversky measure, in Sánchez et al. [53] the authors introduce a feature-based similarity measure based on set theory operations between the ancestor sets of the concepts to be compared. From other standpoint, in Pirró and Seco [45] the authors propose an IC-based reformulation of the Tversky similarity measure that you can see in Table 1. One drawback of the latter measure is that “it treats the similarity between identical concepts as a special case and can give as output negative values” [45, p. 619]. In order to overcome this drawback, in Pirró and Euzenat [44], the authors introduce another reformulation of the Tversky similarity measure, called FaITH, which is shown in Table 1.

In Meng and Gu [31], the authors introduce the IC-based semantic similarity shown in Table 1, which is defined by an exponential-like transformation of the classic Lin measure.

In Lastra-Díaz and García-Serrano [26], we introduce a new family of IC-based similarity measures derived from the Jiang–Conrath distance, which is based on the observation that the conversion of the Jiang–Conrath distance into a similarity measure is an unknown and likely not linear transformation. To bridge this gap, we propose the similarity measure called cosine-normalized Jiang–Conrath (*cosJ&C*) that is shown in Table 1. The *cosJ&C* measure is based on the normalization of the Jiang–Conrath distance through a normalized exponential-like function. In this latter work, we also introduce two hybrid IC-based similarity measures called *weighted J&C* and *cosine-normalized weighted J&C similarity measures*. These latter measures rely on the computation of the length of the shortest path between concepts on a weighted taxonomy, whose edge weights are defined as the information content of the conditional probabilities between child and parent concepts. In our aforementioned work, we carried-out a benchmark with 17 ontology-based similarity measures based on a common code implementation, the winner measures being the *cosJ&C* similarity, the $sim_{Taieb}(c_1, c_2)$ similarity measure proposed in Hadj Taieb et al. [15], and the Meng and Gu [31] measure shown in Table 1. In these benchmarks, we evaluated most of hybrid IC-based similarity measures reported in the literature, such as the pioneering work in Li et al. [29], the measures proposed in Zhou et al. [64], Meng et al. [33] and Gao et al. [12], and our two aforementioned measures called *wJ&Csim* and *coswJ&Csim*. Despite most of hybrid IC-based measures obtained rivaling results as regards the state of the art,

Table 2
State-of-the-art Information Content models evaluated in this work.

IC models	Definition
Resnik [48]	$IC_{Resnik} = -\log_2(\hat{p}(c_i))$, $\hat{p}(c_i) = \frac{f(c_i)}{N} = \frac{f(c_i)}{f(T)}$ $f(c_i) = TF(c_i) + IF(c_i) = TF(c_i) + \sum_{\forall c_j c_i \in LA(c_j)} f(c_j)$
Seco et al.	$IC_{Seco}(c) = 1 - \frac{\log(\text{Hypo}(c)+1)}{\log(\text{max_nodes})}$
Zhou et al.	$IC_{Zhou}(c) = k \left(1 - \frac{\log(\text{Hypo}(c) +1)}{\log(\text{max_nodes})} \right) + (1-k) \frac{\log(\text{depth}(c))}{\log(\text{depth}_{max})}$, $k' = \frac{1}{2}$ (default)
Sebti and Barfroush	$IC_{Sebti}(c) = -\log_2(p)$, p identical to CondProbUniform
Sánchez et al. [52]	$IC_{Sánchez2011}(c) = -\log_2 \left(\frac{\text{Leaves}(c)}{\text{maxLeaves}+1} \right)$
Sánchez et al. [53]	$IC_{Sánchez2012}(c) = -\log_2 \left(\frac{\text{commonness}(c)}{\text{commonness}(\text{root})} \right)$ $\begin{cases} \text{commonness}(c) = \frac{1}{ \text{Subsumers}(c) } & c \text{ leaf} \\ \text{commonness}(c) = \sum_{\forall l l \text{ is leaf and } l \prec c} \text{commonness}(l) & c \text{ not leaf} \end{cases}$
Harispe	$IC_{Harispe}(c) = -\log_2 \left(\frac{\text{Leaves}(c)+1}{\text{Subsumers}(c)} \right)$
Meng et al.	$IC_{Meng}(c) = \frac{\log(\text{depth}(c))}{\log(\text{depth}_{max})} \times \left(1 - \frac{\log \left(1 + \sum_{a \in \text{Hypo}(c)} \frac{1}{\text{depth}(a)} \right)}{\log(\text{Node}_{max})} \right)$
Yuan et al.	$IC_{Yuan}(c) = f_{\text{depth}}(c)(1 - f_{\text{leaves}}(c)) + f_{\text{hyper}}(c)$ $\begin{cases} f_{\text{depth}}(c) = \frac{\log(\text{depth}(c))}{\log(\text{depth}_{max})} \\ f_{\text{leaves}}(c) = \frac{\log(\text{Leaves}(c)+1)}{\log(\text{Leaves}_{max}+1)} \\ f_{\text{hyper}}(c) = \frac{\log(\text{Hyper}(c)+1)}{\log(\text{Node}_{max})} \end{cases}$
Hadj Taieb et al.	$IC(c) = \left(\sum_{a \in \text{HyperInc}(c)} \text{Score}(a) \right) \times \text{AvgDepth}(c)$ $\text{AvgDepth}(c) = \frac{1}{ \text{HyperInc}(c) } \times \sum_{c' \in \text{HyperInc}(c)} \text{depth}(c')$ $\text{Score}(c) = \left(\sum_{c' \in \text{DirectHyper}(c)} \frac{\text{depth}(c')}{ \text{HypInc}(c') } \right) \times \text{HypInc}(c) $ $\text{HypInc}(c) = \{a \in C \mid a \leq c\}$ $\text{HyperInc}(c) = \{a \in C \mid c \leq a\}$

they did not outperform the *cosJ&C* similarity measure, or the Meng and Gu [31] and Hadj Taieb et al. measures. The hybrid *coswJ&C* similarity measure obtains the highest correlation values in the Rubenstein and Goodenough (RG65) dataset, however, this measure, as well as the rest of hybrid measures, did not obtain convincing results that justify their high computational cost. For this reason, we refute the practical use of any current hybrid IC-based measures, unless they are able to outperform other less complex measures convincingly, thus, we discard these hybrid IC-based measures from the experiments herein.

In Hadj Taieb et al. [15] the authors introduce the intrinsic similarity shown in Table 1, which is based on a new way of computing the contribution of the hyponym set of a concept. Despite this measure not being based on an IC model, it is closely related to the Seco et al. IC model and obtains state-of-the-art results in our aforementioned work. Thus, in order to offer a complete image of the state of the art in ontology-based similarity measures, we include the Hadj Taieb et al. measure in our benchmarks.

3. Related work on information content models

In Resnik [48], the author introduces the most broadly accepted corpus-based IC model for the evaluation of semantic similarity tasks, which is shown in Table 2. The Resnik method is based on the estimation of the concept probabilities through the frequency counting of concept occurrences in a training corpus. Each occurrence in the corpus of a word contained in WordNet is counted as an occurrence of all its subsumed concepts. In [41, p.34], Pedersen describes the Resnik frequency counting method used to build the WordNet-based frequency files used in our experiments, Pedersen [39], as well as the corpus-based IC models evaluated in his paper series on similarity measures in WordNet. Following the notation of Pedersen to define the IC_{Resnik} model, each concept frequency $f(c_i)$ is defined as the sum of the term-frequency (TF) occurrences of the concept c_i , plus the inherited frequency (IF) of each subsumed child concept. The estimated probability $\hat{p}(c_i)$ of

each taxonomic concept $c_i \in C$ is defined as the ratio of the concept frequency to the root frequency, where N is the total number of occurrences of any noun within the corpus and its value matches the frequency of the root concept Γ . This frequency counting does not take into account the word senses, although Resnik suggests that a sense-tagged corpus could be used to improve this issue. In another work Pedersen [40], the authors prove that the IC models derived from a non sense-tagged corpus perform better than the sense-tagged ones. Like most IC models, the Resnik method does not satisfy the axioms for a well-founded IC model described in Section 4, encouraging the proposal of the *CondProbCorpus* IC model in order to complete the proposed family herein.

Encouraged by the drawbacks in the aforementioned corpus-based IC models, in Seco et al. [56] the authors introduce the first known intrinsic IC model in the literature. The core idea of the intrinsic IC models is the computation of the IC values using solely taxonomical features, such as: the density of the descendant or ancestor nodes (hyponym set/hypernym set), the subsumed leaf nodes, or the node depth among others. During the last decade, the development of intrinsic IC models has become one of the mainstays of research in the area. Among the main intrinsic IC models proposed in the literature, we find the works in Zhou et al. [63], Sebti and Barfroush [55], Sánchez et al. [52], Sánchez and Batet [51], Yuan et al. [62], Harispe et al. [17] and Hadj Taieb et al. [14]. In Table 2, we summarize the definition of the state-of-the-art intrinsic IC models implemented in our experiments.

The Seco et al. IC model is based on the idea that the information content is inversely proportional to the number of hyponyms of a concept. Looking at the expression for IC_{Seco} in Table 2, we can appreciate that the model is carrying out some type of estimation of the concept probabilities, such as the ratio of the logarithms of hyponyms and the total number of concepts. The Seco et al. IC model assumes, in an underlying way, a uniform occurrence probability for every concept when it is measuring the set of hyponyms. This method does not satisfy the structure axioms introduced herein, moreover the IC values in the leaf concepts are artificially forced to 1. However, this IC model is closely related to the *CondProbHypo* model, in which we use the hyponym ratio between parent and children concepts to estimate their conditional probabilities. In addition, we can prove, taking the limits on the number of total concepts, that the *CondProbHypo* model could be interpreted as a normalization of the Seco et al. IC model that satisfies the probabilistic structure axioms introduced in Section 4.1, when the base ontology is tree-like. Specifically, this normalization is responsible for the better result of *CondProbHypo* on some similarity measures closely related to the conditional probability notion, such as the Jiang–Conrath distance.

In Zhou et al. [63], the authors note that the Seco et al. IC model does not consider the depth of the concepts, thus, two concepts with an equal number of hyponyms, but very different depths, can produce similar IC values. This fact contradicts the expected behavior of the lowest depth concepts, where we expect, according to the IC hypothesis, that the more abstract ones demonstrate a lower information content. Encouraged by this drawback, Zhou et al. introduce an intrinsic IC model, denoted herein as IC_{Zhou} , which is based on a linear combination of the Seco et al. IC model with an estimation of the information content based on the depth of the concepts on the taxonomy. Like the IC model in Seco et al. [56], the Zhou et al. IC model does not consider the structure axioms related to the conditional probabilities.

In Sebti and Barfroush [55], the authors introduce a naive intrinsic IC model which assigns uniform conditional probabilities between each concept and its children, then the probabilities and IC values of the taxonomy nodes are recovered by applying the same iterative algorithm that we use in our IC models. Although the authors do not explain how to compute the model for ontologies

with multiple inheritance as we have done, it is obvious that they must be aware of this case. Their IC model matches the simplest model of our family, called *CondProbUniform*, thus, we consider this model as our direct predecessor. Unlike this work, they do not propose a general structure-preservation framework as introduced herein. Despite developing our IC models in an independent manner, we consider the Sebti et al. IC model as the first well-founded IC model based on the notion of conditional probability, according to the general framework of well-founded IC models proposed herein.

In Blanchard et al. [5], the authors introduce a collection of intrinsic IC models based on different hypotheses to estimate the concept probabilities. Two of their methods are related to our *CondProbUniform* and *CondProbLeaves* IC models, however, like other models, they do not consider the structure axioms defined in our family of well-founded IC models. In order to evaluate their models, the authors compute the correlation between each IC model and one corpus-based IC model. The paper only reports the correlation values for two intrinsic IC models based on the node depths and subsumed leaves ratio, reporting respectively an approximated correlation values of 0.3 and 0.5 in WordNet. The comparison method introduced by the authors is very appropriate for evaluating directly the fitting quality of the intrinsic IC models as regards the corpus-based ones, in order to estimate how well an intrinsic IC model approximates a corpus-based model. Indeed, in the light of the findings in Lastra-Díaz and García-Serrano [26] and this work, we have considered following a similar approach into a future deeper study between intrinsic and corpus-based IC models. Despite the proposed IC models deserving to be evaluated, the work does not report any result based on any standard similarity benchmark on the family of IC-based measures, such as we do herein. Therefore, we have discarded to evaluate them in our experiments due to the impossibility of replicating and comparing their results.

In Sánchez et al. [52], the authors argue that the hyponym set of a concept includes many abstract concepts which rarely occur in any corpus, which follows that the use of the hyponyms set, such as that in Seco et al. [56], is not appropriate to estimate the IC values. Encouraged by their observation, Sánchez et al. propose an IC model based on a ratio of leaves and subsumers of a concept with regard to the total number of concepts. The use of leaves count makes the model less dependent of the number of inner (abstract) concepts of any taxonomy. Despite the argument against the hyponyms being plausible, and well-founded, we do not think that it is completely true in daily language. Like the other aforementioned models, this model does not verify the structure axioms introduced in this work.

In addition to the arguments of Sánchez et al. [52] endorsing the use of the leaves as features to compute the IC model, we provide here other significant argument to endorse this idea, which is closely connected with the same structuralist design-principles followed herein. The canonical definition of a discrete probability space in any taxonomy, such as that introduced in Section 4.1, starts with the assignment of a probability function based on a partition of the entire sample space defined by the leaf concepts. Therefore, the true probability of any subset of the sample space is simply the sum of the probabilities of its subsumed leaf concepts. From this point of view, we would expect that the leaf-based IC model, called *CondProbLeaves*, or the Sánchez et al. [52] model, would have a clear advantage over the rest of methods.

Following the ideas of Sánchez et al. [52], Harispe introduced a minor variant of this model, as shown in Table 2. The new IC model is called *Harispe2012* in the implementation code of the Semantic Measures Library Harispe et al. [17]. Despite this IC model has not published in the literature, we have implemented it in our experiments for the sake of completeness.

In Sánchez and Batet [51], the authors introduce an intrinsic IC model based on a notion called commonness, which derives from a

different use of the leaves and subsumer sets associated to any concept. Their model tries to capture the occurrence probability of each leaf, such that the leaves with lower occurrence probability have a lower information content. The probability of every leaf concept is approximated by the inverse of the number of subsumer concepts. This model does not consider the structure axioms introduced herein.

In Meng et al. [32], the authors introduce the intrinsic IC model shown in Table 2, which is based on the following taxonomic features: (1) the depth of the concept, (2) the depth of the hyponyms of the concept, and (3) the number of hyponyms. The model is defined by the product of two different estimations of the IC value. The first one matches the depth ratio introduced in Zhou et al. [63], while the second one is similar in structure to the hyponym-based Seco et al. model, but unlike the number of hyponyms, Meng et al. use the inverse of the depth values associated to the hyponyms. The motivation of this work follows the ideas behind the use of the hyponyms and depth features in Zhou et al. [63], but Meng et al. also consider some new cases: (1) two concepts with the same number of leaves, but a different number of hyponyms should have a different IC value, and (2) two concepts with the same number of hyponyms, but at different depths, should also have different IC values. The core idea of this work is to take into account the depth of each hyponym in its contribution to the IC estimation. Unlike the Zhou et al. model, which uses a linear combination of IC estimation factors, this IC model combines two independent estimations of the IC values through a product. Finally, this model is the first one to introduce the inverse of depth as a taxonomic feature for the implicit estimation of the occurrence probability for each concept.

In Yuan et al. [62], the authors introduce an intrinsic IC model which could be categorized, together with the models of Zhou et al. [63] and Meng et al. [32], in a subfamily of intrinsic IC models that combines taxonomic features introduced in other models, or new variants, using linear combinations or products of them. This model combines three independent factors in a closed formula, from among which we have the depth ratio introduced in Zhou et al. [63] and also used in Meng et al. [32], as well as two new ratio factors based on leaves and hypernyms. We note that the parameters of the logarithm in the formulas in Table 2 includes the sum of a 1 factor because the *leaves(c)* and *hypo(c)* functions do not count the input concept.

Finally, in Hadj Taieb et al. [14], the authors introduce another intrinsic IC model shown in Table 2, whose formulation relies on a new method of evaluating the contribution of the hyponym set. The model is based on the product of the average depth ratio of the concept with regard to the number of hyponyms, and a weighted linear combination of the average depth of the concepts in the hyponym sets. The core contribution of the model is the definition and use of the average depth for the hyponym sets.

In summary, the review of the state of the art shows that a broad set of taxonomic features has been successfully proposed to compute the IC models. However, most of intrinsic IC models neither consider the structure axioms derived from the conditional probability, nor the cognitive similarity that we introduce in this work.

4. The family of well-founded IC models

In this section, we introduce the new family of well-founded IC models based on the conditional probability notion between child and parent concepts. The new IC models are based on the estimation of the conditional probabilities between child and parent concepts, which are edge-valued functions. Therefore, the proposed method for the design of new IC models is edge-based, unlike most previous models in the literature that use some type of node-based computational method, with the exception of Sebti and Barfroush [55]. Despite only proposing three different methods for the intrinsic

estimation of the conditional probabilities herein, we define an open framework for the design of new IC models based on the proposal of alternative methods for their estimation. Second, we introduce the use of two cognitive-based scaling functions (sigmoid and exponential) as a means of encoding a notion of cognitive similarity, as suggested by Gärdenfors in [13, Section 2.8], according to some results obtained in cognitive psychology.

4.1. Preliminary concepts and notation

For the sake of clarity, we use the lowercase letter p to denote a concept-valued probability function in a set of concepts C . On the other hand, the uppercase P is reserved to denote a probability measure, which is a set-valued function in the power set of the sample space. Finally, the conditional probability functions between concepts are denoted in lowercase by $p(c_i|c_j)$.

All the IC models proposed herein share the same computational structure, defined by the following three steps: (1) estimation of the edge-based conditional probabilities $p(c_i|c_j)$, (2) recovery of the concept-valued probability density function $p(c_i)$, and (3) computation of the node-based IC values using the standard definition $IC(c_i) = -\log_2(p(c_i))$. The only difference between the IC models is the method used to estimate the conditional probabilities. We call the new IC models *well-founded* because they are designed from first principles in order to satisfy the structural relationships of a discrete probability space and an information content model defined on this space.

In Jiang and Conrath [23], the authors prove that their semantic distance $d_{J&C}(c_1, c_2)$ is equivalent to the length of the shortest path between concepts c_1 and c_2 over a weighted graph derived from the taxonomy, and the edge weights are defined by (2). Despite the authors claiming that their distance is a metric on any type of taxonomy, in Orum and Joslyn [36] the authors prove that it is only true for the tree-like taxonomies, not for general taxonomies with multiple inheritance.

Every taxonomy $\mathcal{C} = (C, \leq_c, \Gamma)$ induces a graph $G = (E, V)$ in the usual manner, where every concept is a vertex of the graph, it means $V = C$, and there is an edge between each concept c_i and its direct parents, also called the lowest ancestors of c_i and denoted as $LA(c_i)$. The IC-based weighting function (2) allows us to introduce a shift of paradigm for the definition of the IC models, we move from a node-based IC computation model to an edge-based model.

$$\begin{aligned} w : E &\rightarrow \mathbb{R} \\ w(e_{ij}) &= -\log_2(p(c_i|c_j)) = IC(c_i) - IC(c_j) \\ E &= \{(c_i, c_j) \subset C \times C \mid c_j \in LA(c_i)\} \end{aligned} \quad (2)$$

Formally, a probability space is a triplet (Ω, \mathcal{F}, P) , where Ω is a non-empty set, called the space of outcomes or samples, \mathcal{F} is a σ -algebra that defines the collection of all possible events, where every event is defined as a subset of Ω , and finally, $P : \mathcal{F} \rightarrow \mathbb{R}$ is a probability measure. The formal definitions of the probability measures and probability spaces can be consulted in [2, Section 1.2].

Definition 1 (*Probability measure*). Given any non-empty set Ω and a collection \mathcal{F} of subsets on Ω , such that \mathcal{F} is a σ -algebra, then a set-valued function $P : \mathcal{F} \rightarrow \mathbb{R}$ is a probability measure if it satisfies the following axioms:

1. $0 \leq P(A) \leq 1, \forall A \in \mathcal{F}$.
2. $P(\Omega) = 1, P(\emptyset) = 0$.
3. If $A = \{A_1, A_2, \dots, A_n\}$ is a family of disjoint subsets of \mathcal{F} , such that $\forall A_i, A_j \in A \Rightarrow A_i \cap A_j = \emptyset$, then:

$$P\left(\bigcup_{i \in I} A_i\right) = \sum_{i \in I} P(A_i)$$

If the space Ω is a countable set, and \mathcal{F} is the power set of the sample space Ω , denoted as 2^Ω , the triplet (Ω, \mathcal{F}, P) is called a discrete probability space [2, Section 4.2]. In our case, the space of samples Ω is discrete and is defined by the root concept Γ , such that $\Omega := \Gamma$, and the set \mathcal{F} is only the power set of the root concept Γ . Here, we are defining the root concept Γ as the universal set of the taxonomy, which follows that $\Gamma := C$. We note that in the last statement we are abusing the notation, because Γ is used to denote the root element of C and the sample space Ω at the same time.

We recall that the power set 2^Ω of any set Ω is a complete lattice when the inclusion relation \subseteq between subsets in Ω is used as an order relation. This fact is closely related to the relationship between the Jiang–Conrath distance and some types of metrics on lattices, such as we note in Lastra-Díaz [24], and is detailed in a work on the metric properties of the Jiang–Conrath distance Orum and Joslyn [36].

Given a taxonomy $\mathcal{C} = (C, \leq_C, \Gamma)$, where $L_C = \{c_k \in C \mid \nexists c_i \neq c_k, c_i \leq_C c_k\}$ is the set of leaves of the taxonomy, we can define a discrete probability space (Γ, \mathcal{F}, P) on \mathcal{C} in the canonical manner as follows: (1) we define the root concept Γ as the universal sample space, (2) we define the set of leaf concepts $L_C \subset C$ as the partition of disjoint sets of the sample space, such that $\Gamma := L_C$ by definition, (3) we define \mathcal{F} as the power set on C , such that $c_i \subseteq c_j \iff c_i \leq_C c_j$, and finally, (4) we define a set-valued function $P: \mathcal{F} \rightarrow [0, 1]$ using a normalized leaf-valued function $p(c_k)$.

The triplet (Γ, \mathcal{F}, P) , as defined above, is a well-founded discrete probability space, a fact that we formalize in Proposition 1, whose formal proof is omitted through lack of space. In order to get a well-founded probability space on any taxonomy, and derive the new family of intrinsic IC models from it. Below we provide a method to define any well-founded IC model based solely on the estimation of the conditional probabilities, which constitutes the core idea of this work.

Definition 2 (well-founded IC model). Given a taxonomy of concepts $\mathcal{C} = (C, \leq_C, \Gamma)$, and an IC model defined by the function $IC: C \rightarrow \mathbb{R}^+ \cup \{0\}$, we call it a well-founded IC model if it can be written as $IC(c) = -\log_2(p(c))$ where $p(c)$ is a concept-valued function defined by (4), and the functions $p(c_i|c_j)$ are the conditional probabilities between any child concept c_i and its parent concepts c_j , which satisfy the edge-based property in (3).

(1) *Edge-based axiom.* The sum of conditional probabilities $p(c_i|c_j)$ of the children nodes c_i on any parent c_j node must be equal to 1, as defined in Eq. (3), where $LA(c_i)$ denotes the set of lowest ancestors (direct parents) of any concept c_i .

$$\sum_{\forall c_j|c_j \in LA(c_i)} p(c_i|c_j) = 1 \quad (3)$$

(2) *Node-based axiom.* The probability $p(c_i)$ for each node c_i must be equal to the integration of the probabilities throughout the graph, starting from the root node, as defined in Eq. (4).

$$p: C \rightarrow [0, 1] \subset \mathbb{R}$$

$$p(c_i) = \begin{cases} 1, & c_i = \Gamma \\ \sum_{\forall c_j \in LA(c_i)} p(c_j)p(c_i|c_j), & c_i \neq \Gamma \end{cases} \quad (4)$$

(3) *Leaf-based axiom.* The probabilities of the leaf concepts sum 1.

$$\sum_{c_k \in L_C} p(c_k) = 1 \quad (5)$$

The axioms (1) and (2) above allow us to define a family of well-founded intrinsic IC models based on the estimation of the

Table 3

Family of well-founded intrinsic and corpus-based IC models proposed in this work. For all the IC models, $P(c_i)$ is recovered using the probability recovery Algorithm 1.

IC models	Definition
CondProbHypo	$IC_{CPHypo}(c_i) = -\log_2(p_{Hypo}(c_i))$ $p_{Hypo}(c_i c_j) = \frac{ Hypo(c_i) +1}{\sum_{\forall c_k c_j \in LA(c_k)} (Hypo(c_k) +1)}$
CondProbUniform	$IC_{CPUni}(c_i) = -\log_2(p_{Uniform}(c_i))$ $p_{Uniform}(c_i c_j) = \frac{1}{ children(c_j) }$
CondProbLeaves	$IC_{CPLea}(c_i) = -\log_2(p_{Leaves}(c_i))$ $p_{Leaves}(c_i c_j) = \frac{ Leaves(c_i) +1}{\sum_{\forall c_k c_j \in LA(c_k)} (Leaves(c_k) +1)}$
CondProbCorpus	$IC_{CondProbCorpus}(c_i) = -\log_2(p(c_i))$ $p(c_i) = \begin{cases} 1, & c_i = \Gamma \\ \sum_{\forall c_j \in LA(c_i)} p(c_j)p_{corpus}(c_i c_j), & c_i \neq \Gamma \end{cases}$ $p_{corpus}(c_i c_j) = \frac{\max\{f(c_i)\}}{\sum_{\forall c_k c_j \in LA(c_k)} \max\{f(c_k)\}}$
CondProbLogistic	$IC_{CPLog}(c_i) = -\log_2(p_{Log}(c_i))$ $p_{Log}(c_i c_j) = \varphi_1 \circ p_{Hypo}(c_i c_j)$ $\varphi_1(x: k) = \frac{1}{1+e^{-k(\frac{x}{2})}}, k^* = 8$
CondProbCosine	$IC_{CPCos}(c_i) = -\log_2(p_{Cos}(c_i))$ $p_{Cos}(c_i c_j) = \varphi_c \circ p_{Hypo}(c_i c_j)$ $\varphi_c(x) = 1 - \cos(\frac{\pi}{2}x)$

conditional probabilities $p(c_i|c_j)$ for each edge of the taxonomy, such as is shown in Table 3. In Proposition 1, we show that given a taxonomy (C, \leq_C, Γ) , the leaf-based axiom (3) is a sufficient condition to get a well-founded probability space. In addition, we show in Proposition 2 that axioms (1) and (2) of a well-founded IC model are sufficient conditions to build a leaf-valued function $p: L_C \subset C \rightarrow [0, 1]$ that satisfies the IC model axiom (3). Thus, this last proposition proves that any well-founded IC model induces a well-founded probability space on any base taxonomy, and the whole system is supported by the structures derived from the conditional probabilities. We omit all the proofs herein by lack of space.

Proposition 1. Be a taxonomy $\mathcal{C} = (C, \leq_C, \Gamma)$ defined by a partially ordered set (C, \leq_C) with a distinguished supreme element Γ , called the root, and L_C the set of leaves in C . If a set-valued positive function P is defined from the leaf-valued function p as follows:

$$(1) \quad P: 2^\Gamma \rightarrow [0, 1]$$

$$P(A) = \sum_{c_k \in L_C \cap A} p(c_k)$$

$$(2) \quad p: L_C \subset C \rightarrow [0, 1]$$

$$\sum_{c_k \in L_C} p(c_k) = 1$$

then the following facts are satisfied: (1) P is a probability measure, and (2) the triplet $(\Gamma, 2^\Gamma, P)$ is a probability space.

Proposition 2. Let a taxonomy $\mathcal{C} = (C, \leq_C, \Gamma)$ and L_C be the set of leaves in C . Given a concept-valued function p defined by

$$p: C \rightarrow [0, 1]$$

$$p(c_i) = \begin{cases} 1, & \text{if } c_i = \Gamma \\ \sum_{\forall c_j \in LA_C(c_i)} p(c_j)p(c_i|c_j), & \text{otherwise} \end{cases}$$

then $P(L_C) = 1$, as given below:

$$P(L_C) = \sum_{c_k \in L_C} p(c_k) = 1$$

4.2. Construction of the well-founded IC models

In this section, we introduce three new intrinsic IC models based on different methods to estimate the conditional probabili-

ties, called *CondProbHypo*, *CondProbUniform* and *CondProbLeaves*, which are disclosed in Lastra-Díaz and García-Serrano [25]. In addition, we introduce two additional intrinsic IC models, called *CondProbLogistic* and *CondProbCosine*, which rely on the composition of the hyponym-based conditional probability estimation with two different cognitive-based similarity non-linear functions. Finally, with the aim of bridging the same structure gap observed in the corpus-based Resnik IC model, we introduce the *CondProbCorpus* IC model to complete the family of well-founded IC models.

As we have already outlined in Section 4.1, the computation of our IC models is carried-out in three steps: (1) estimation of the conditional probabilities $p(c_i|c_j)$, (2) recovery of the probability $p(c_i)$ from the taxonomy structure and the $p(c_i|c_j)$ values, and (3) computation of the IC values using Eq. (1). From axiom 2 we can derive a direct algorithm for step (2), as defined below. **Algorithm 1** works on any type of taxonomy, and satisfies the structure axioms 2 and 3 in definition 2, whenever the conditional probabilities $p(c_i|c_j)$ satisfy axiom 1, as is proven in proposition 2. This algorithm is graphically explained for tree-like taxonomies in Sebti and Barfroush [55], although the authors do not follow a generalized and formal structure-preserving approach, as we do herein.

Algorithm 1 (*Probability and IC recovery*). This algorithm takes as inputs: (1) a taxonomy $\mathcal{C} = (\mathcal{C}, \leq_{\mathcal{C}}, \Gamma)$, and (2) a set of conditional probabilities $p(c_i|c_j)$ for each edge within taxonomy. Then the algorithm computes the probability and IC value for each concept $c_i \in \mathcal{C}$ as follows: (1) the probability computation method must build a total ordering of the concept in the taxonomy, which is defined by an ordered list of concepts, such that every concept is in a subsequent position to every one of its parent concepts; (2) the method assigns a value of “1” to the probability of the root concept Γ ; (3) the algorithm traverses the concept nodes according to the previously built total ordering, then it computes the probability $p(c_i)$ of every child concept as the sum on each parent of the product of the parent probability through the conditional probabilities $p(c_i|c_j)$, as defined by formula (4) in Definition 2 above; and finally (4) using the probabilities $p(c_i)$, we compute the IC values as $IC(c_i) = -\log_2(p(c_i))$.

In Table 3, we summarize the six new IC models proposed herein. The first four IC models correspond to four different methods of estimating the conditional probabilities $p(c_i|c_j)$ between any concept c_i and its parent concepts $c_j \in LA(c_i)$. The models *CondProbHypo*, *CondProbUniform*, *CondProbLeaves* and *CondProbCorpus* satisfy the structure axioms introduced in the previous section, and all them share the same computational and algebraic structure. However, the first three models are intrinsic and the last one relies on corpus statistics. These models are computed in three steps as follows: (1) computation of the conditional probabilities $p(c_i|c_j)$, (2) recovery of the probabilities $p(c_i)$ from $p(c_i|c_j)$ using Algorithm 1, and (3) computation of the IC values using Eq. (1). For the formulas in Table 3, *Hypo*(c_i) and *Leaves*(c_i) denote respectively the set of subsumed concepts and leaf concepts for any concept $c_i \in \mathcal{C}$, without including the base concept c_i .

In Table 3, the *CondProbLogistic* and *CondProbCosine* IC models represent two cognitive-based IC models, whose kernel functions are shown in Fig. 1. We consider the conditional probabilities as a linear measure of the degree of cognitive similarity between the children and parent concepts in the taxonomy, because it is the right place to integrate this notion of cognitive similarity within the new IC models. Following the cognitive ideas suggested in [13, Section 2.8], we define the cognitive similarity functions φ_l and φ_c in the equations (6) and (7).

$$\varphi_l/\varphi_c : [0, 1] \subset \mathbb{R} \rightarrow [0, 1] \subset \mathbb{R}$$

$$\varphi_l(x : k) = \frac{1}{1 + e^{-k(x-\frac{1}{2})}} \quad (6)$$

$$\varphi_c(x) = 1 - \cos\left(\frac{\pi}{2}x\right) \quad (7)$$

Both functions, φ_l and φ_c , are combined with the *CondProbHypo* model, because it obtains the best results from among the non-cognitive models. Unlike the other non-cognitive IC models, these models are computed in four steps: (1) the computation of the conditional probabilities $p_{Hypo}(c_i|c_j)$, (2) the computation of the cognitive conditional probabilities $p_{Log}(c_i|c_j)$, or $p_{Cos}(c_i|c_j)$, (3) the recovery of the probabilities $p(c_i)$ from $p_{Log}(c_i|c_j)$, or $p_{Cos}(c_i|c_j)$ using the recovery algorithm 1, and (4) the computation of the IC values using Eq. (1).

The function φ_l is a translated logistic function whose sigmoid shape is defined by the constant k , as shown in Fig. 1. According to our experiments, our preferred default value is $k = 8$. On the other hand, the φ_c corresponds to a scaling and translation of the cosine function to obtain a normalized exponential-like function. The sigmoid function φ_l is explicitly modeling the notion of cognitive similarity suggested by Gärdenfors, it is an instance of a concept being more quickly perceived as belonging to another category, as the distance from the prototype of the category increases.

Inspired by the successful exponential-type scaling defined in the similarity measure of Meng and Gu [31], we wanted to study a normalized exponential-shape function to compare its results with the sigmoid function φ_l . For this reason, we defined the cosine-based function φ_c shown in Fig. 1, which only matches the cognitive similarity criterion for the low range of the function, not the top part.

The final cognitive IC models generated by the *CondProbLogistic* and *CondProbCosine* models are encoding a cognitive similarity, not a classic information content measure, although their genesis and derivation follow the IC approach. We also note that the functions $p_{Log}(c_i|c_j)$ and $p_{Cos}(c_i|c_j)$ do not satisfy the edge-based axiom 1 and the leaves-based axiom 3, thus, despite the functions $p_{Log}(c_i)$ and $p_{Cos}(c_i)$ being computed using the probability recovery Algorithm 1, these functions do not allow formal probability measures to be derived, according to the construction for *well-founded IC models* introduced in the last section. This latter fact could look contradictory to the structure-preserving spirit that this work defends, however, this approach has allowed us to find the place to integrate the notion of cognitive similarity into our family of IC models. This possibility, together with the cognitive similarity notion, has led us to create two IC models that rival the state-of-the-art IC models. These encouraging results show that it is a new line of research that deserves to be explored.

5. Evaluation

The goals of the experiments described in this section are as follows: (1) the experimental evaluation of the proposed IC models and their comparison with the state-of-the-art methods, (2) a new and conclusive experimental study on the state of the art in ontology-based similarity measures, (3) the replication of previously reported methods and results, (4) a new comparison between intrinsic and corpus-based IC models, (5) a study into the impact of the IC models on the IC-based similarity measures, (6) a new confirmation of the findings in [26] related to some previous conclusions on the refuted outperformance of the intrinsic IC models over the corpus-based ones, and (7) a new confirmation of the achievements of the family of intrinsic IC models and IC-based similarity measures.

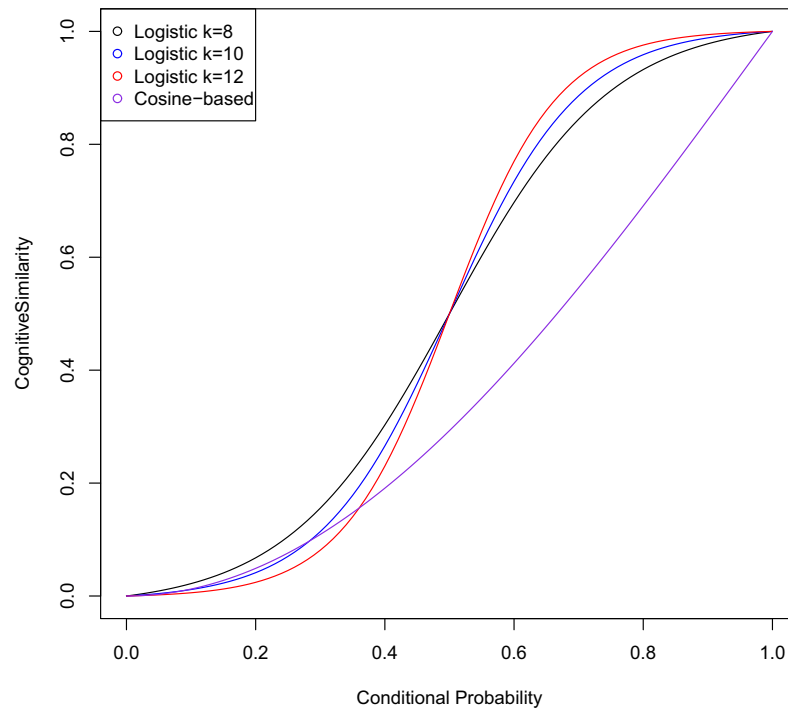


Fig. 1. Cognitive similarity functions used in CondProbLogistic and CondProbCosine.

5.1. Information content models and similarity evaluated measures

In order to compare the new family of IC models in Table 3 with the state-of-the-art models, we implemented all the intrinsic IC models shown in Table 2, as well as all the IC-based similarity measures shown in Table 1. In addition, we implemented the $sim_{T_{aiab}}$ measure shown in Table 1 to obtain a complete image of the state of the art in ontology-based similarity measures. For the $sim_{T_{aiab}}$ measure, the depth is defined by the authors as the longest ascending path length, whilst the rest of measures and IC models consider the depth as the length of shortest ascending path from each concept to the root. For the Zhou et al. IC model, the authors define the depth starting at 1 for the root concept. Our interest in the replication of previous methods and results lead us to implement all IC models and similarity measures evaluated herein. However, we refer the readers working on applications to the SML library, Harispe et al. [17]. In the context of our research, we developed a new data structure for large ontologies that overcome the SML library in computation time and memory use, which we plan to introduce in a future publication.

To the best of our knowledge, we evaluate all the intrinsic IC models reported in the literature, except those introduced in Blanchard et al. [5]. We recall that the Sebti et al. IC model is identical to the *CondProbUniform* IC model. Likewise, we evaluate the most relevant ontology-based similarity measures in WordNet in accordance with to our recent benchmarks in Lastra-Díaz and García-Serrano [26]. Therefore, the experiments herein, together with our aforementioned work, are the largest experimental survey of IC-based models and ontology-based similarity measures reported up to date.

In order to compare the intrinsic and corpus-based IC models, we used two unexplored Wordnet-based frequency files from the family of “add1” frequency files in Pedersen [39]. These files are as follows: (1) “ic-semcorraw-add1.dat”, and (2) “ic-treebank-ad d1.dat”. The frequency files are based on WordNet 3.0 and the Resnik method, as described in [41, p.34]. The selected files encode the corpus-based IC models obtaining the best performance in the benchmarks in our aforementioned work, and we use them to

build four IC models based on the Resnik method and the new *CondProbCorpus* IC model.

5.2. Experimental setup

For the experiments, we use the noun database of Wordnet 3.0 and five known word similarity benchmarks: (1) the RG65 dataset made up of 65 word pairs Rubenstein and Goodenough [49]; (2) the MC28 dataset introduced in Miller and Charles [35], which is made up of a subset of 28 word pairs in the RG65 dataset; (3) the Agirre203 dataset introduced in Agirre et al. [1], which is made up of 203 word pairs and it is a subset of the WordSim-353 dataset; (4) a recent replication of the RG65 dataset called $P\&S_{full}$, which is introduced in Pirró [43]; and (5) the SimLex-999 dataset introduced in Hill et al. [20], which is the largest and most recent word similarity benchmark in the literature. All the datasets are defined by a collection of word pairs contained in the noun database of Wordnet, together with a human judgement of its degree of similarity.

Some preprocessing was necessary for the Agirre203 and SimLex-999 datasets to carry out the experiments. For the Agirre203 dataset, it was necessary to remove two word pairs containing verbs not present in the noun database of Wordnet 3.0, such as the pairs (*drink, eat*) and (*stock, live*). In addition, it was also necessary to change the term “media” for “medium”, and “children” for “child”, because these terms do not appear directly in noun database. For this reason, we only used 201 nouns instead of 203, thus, this subset is called hereafter Agirre201. In the case of SimLex-999, it contains 666 nouns, but the word “August” is not included as synset in WordNet 3.0, thus, we only used 665 nouns from the SimLex-999 dataset, and this subset is called hereafter SimLex665.

5.3. Evaluation metrics

As evaluation metrics, we use the Pearson correlation factor, denoted by r in Eq. (8), and the Spearman rank correlation factor denoted by ρ in Eq. (9). The Pearson correlation is invariant as

regard any scaling, translation and rotation of the data, that is any Euclidean similarity. On the other hand, the Spearman correlation factor is rank invariant, what means that it holds the same value for any monotone data transformation. From their invariance properties, it follows that the Pearson correlation encodes the differences in the ratios between different data components X_i and Y_i , whilst the Spearman correlation encodes the differences in the relative ranking between data components. The Pearson correlation compares data vectors in the n -dimensional Euclidean space, whilst the Spearman correlation compares the ranking of the data components, thus, a Pearson correlation of 1 implies a Spearman correlation of 1, but not the opposite.

$$r = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}} \quad (8)$$

$$\rho = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)}, \quad d_i = (x_i - y_i) \quad (9)$$

Given n samples of two independent random variables X and Y , the Pearson correlation computes a value that matches the normalized dot product between the two vectors representing the samples of each random variable. In Eq. (8), X_i are the correlation values reported by any measure for each word pair, and the Y_i values correspond to the human judgements. In Eq. (9), x_i and y_i are respectively the ranking position of the X_i and Y_i values, with the following special case: if some data values share the same rank, their rank is set to the average value that they would have if their rank were different.

The evaluation methodology is based on the performance of the IC models in some word similarity tasks, thus, the evaluation is task-oriented. In contrast, the approach followed in Blanchard et al. [5] measures the fitting quality of the intrinsic IC models with regards to the corpus-based ones. The Blanchard et al. evaluation method uses the correlation between IC models as a quality measure of the underlying function approximation problem, approach that we plan to explore in greater detail in a forthcoming study into this problem. The overall performance of the IC-based similarity measures combined with all the IC models is used to define the performance of the IC models, as well as the performance of the similarity measures.

5.4. Polysemic words

In Wordnet, every word has multiple meanings, also called *synsets*, where every meaning defines a particular concept. For this reason, we follow the same approach as the rest of works in the area [53, Section 3.4]. Given two input words, we evaluate the similarity for the Cartesian product of their synset sets, then, we choose the higher similarity result. This approach follows the idea that any human being uses to select the closest meanings from among word pairs to evaluate their degree of similarity. However, another possible approach could be to select the average of the pairwise similarity values for the synset sets.

5.5. Results

In Tables 4–6 we introduce the summary tables with the overall results for the IC models and the similarity measures, whilst the detailed results per dataset are shown in Tables 7–11 in the appendix.

In Table 4 we show the best result obtained for each IC model with any similarity measure on each dataset and the overall average correlation over all the datasets, regardless of the IC-based similarity measure that produces the best value. The table is ordered according to the average Pearson correlation values. On

the other hand, Table 5 shows the average correlation values for each pair (IC model, IC measure) on all the datasets evaluated, which represent the cell-based average of the Tables 7–11. Table 5 allows to the overall performance of each pair (IC model, IC measure) to be evaluated, as well as studying the impact of the IC models on the different measures. In Table 6, we show a summary with the best results obtained for each similarity measure on each dataset, regardless of the IC models.

In order to evaluate the statistical significance of the data, we prefer confidence intervals over the p-value method. For a brief comparison of both contrast hypothesis methods, we refer the reader to the tutorial in du Prel et al. [10]. Figs. 2 and 3 show the confidence intervals for the difference mean between the IC models and IC measures as regards to their baselines. In both cases, the null hypothesis is that the difference mean is 0, thus, both models perform equally. Any IC model or similarity measure will only have a statistically significant difference over the baseline if its interval does not include the zero level line. Fig. 2 shows the confidence intervals for the difference mean of the average Pearson correlation values in Table 5 between each IC model and the baseline defined by the *Resnik_{ic-trb}* IC model. For each IC model (row), we define a random variable by subtracting the IC model baseline, thus, we are studying the randomness of the IC models as regard the similarity measures. In Fig. 2, we have omitted the representation of the *CondProbUniform* and *Hadj Taieb et al.* IC models, because these models obtain significant statistical lower results than the baseline. Following the same approach, Fig. 3 shows the confidence intervals for the difference mean of the average Pearson correlation values in Table 5 between each IC-based similarity measure and the baseline defined by the J&C similarity, when their values are compared for each IC model.

6. Discussion

Table 4 shows that if we solely consider the best results of each IC model on each dataset, the corpus-based *Resnik_{ic-trb}* obtains the highest average Pearson correlation value from among all the IC models evaluated herein. Analyzing the results on each dataset, we conclude the following: (1) the *Resnik_{ic-trb}* IC model combined with the J&C similarity measure obtains the highest Pearson correlation on the MC28 dataset, (2) the Yuan et al. IC model obtains the highest Pearson correlation on the Agirre201, *P&S_{full}* and *SimLex-665* datasets when it is combined with the *FaITH* and *cosJ&C* similarity measures, and (3) the Sánchez et al. [52] IC model obtains the highest Pearson correlation on the RG65 dataset with the *cosJ&C* similarity measure.

On the other hand, the data in Table 5 show that the Seco et al. IC model obtains the higher average overall performance regardless of the IC-based similarity measure used. The Yuan et al., Meng et al. and Sánchez et al. [53] IC models follow the Seco et al. model in the ranking. Most of the *CondProb* IC models obtain rivaling results in the middle, and the *CondProbLogistic* IC models perform slightly worse. The *CondProbUniform* and *Hadj Taieb et al.* IC models obtain the lowest statistical significant results, thus, these IC models are not shown in Fig. 2 and they should be discarded in future studies.

6.1. The statistical significance of the results

Despite the Seco et al. IC model obtaining the highest average performance, a detailed analysis of the level of significance of the data in Table 5 reveals other conclusions. The confidence intervals in Fig. 2 show that there is no a significant statistical difference between the *Resnik_{ic-trb}* IC model (baseline) and the following IC models: Seco et al., Yuan et al., Meng et al.,

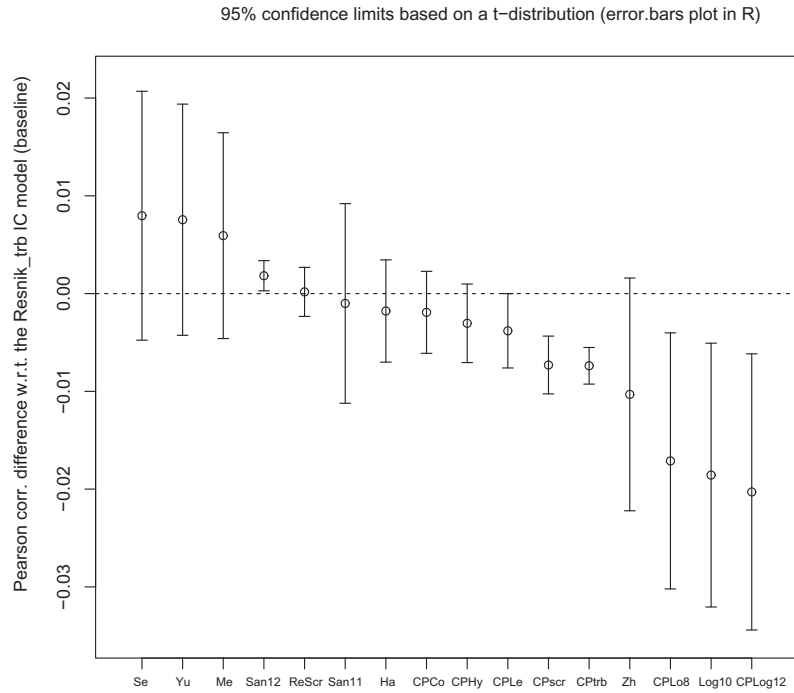


Fig. 2. Mean of the random variable X for the null hypothesis (baseline) defined as $X = Pearson(IC_{model}) - Pearson(Resnik_{trb})$ for each IC-based similarity measure.

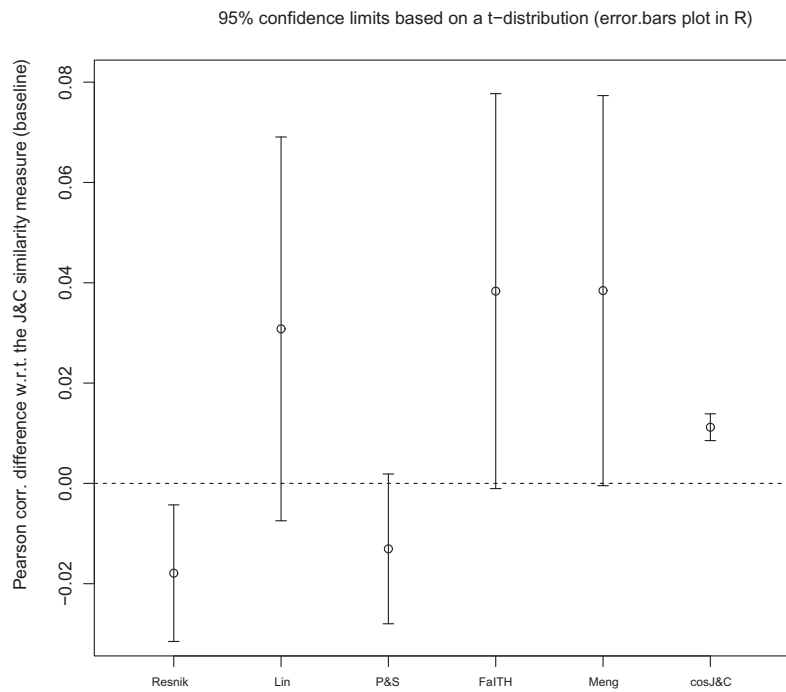


Fig. 3. Mean of the random variable X for the null hypothesis (similarity measure baseline) defined as $X = Pearson(sim_{measure}) - Pearson(sim_{J&C})$ for each IC model.

Resnik_{ic-semcorraw-add1}, Sánchez et al. [52], Harispe, *CondProbCosine*, *CondProbHypo*, and Zhou. This latter fact proves that these intrinsic IC models perform at least as well as the best corpus-based IC model evaluated in our aforementioned work and herein.

On the other hand, Fig. 2 also shows that Sánchez et al. [53] is the only IC model that obtains a statistically significant higher performance than the baseline, regardless of the IC-based similarity measure used. Therefore, the Sánchez et al. [53] IC model is a convincing winner according to the overall average performance and

the statistical evidence obtained, regardless of the selected IC-based similarity measure. This finding implies that in a blind scenario, Sánchez et al. [53] should be the preferred IC model.

Other IC models exhibit a statistically significant lower performance than baselines, such as: *CondProbLeaves*, *CondProbCorpus*, *CondProbLogistic*, *CondProbUniform* and Hadj Taieb et al. IC models. Despite expecting good results for the *CondProbLeaves* IC model through its close relationship with the definition of a probability space, the available statistical evidence refutes this hypothesis.

6.2. Intrinsic versus corpus-based IC models

The statistical evidence proves that a set of intrinsic IC models is a practical alternative to the best performing corpus-based IC models. However, the same evidence allows some previous conclusions on the outperformance of the intrinsic IC models over the corpus-based to be refuted, a finding that we also report in Lastra-Díaz and García-Serrano [26]. Some recent comparisons between intrinsic and corpus-based IC models, such as Sánchez et al. [53] and Yuan et al. [62], are based on the results reported in Patwardhan and Pedersen [38], Pedersen [40] and Pirró [43] for some corpus-based IC models derived from a different set of Wordnet-based frequency files that are used herein. The outperformance of the $Resnik_{ic-trebank-add1}$ IC model in Table 4, and the statistical evidence shown in Fig. 2, conclusively confirm that most of state-of-the-art IC models do not outperform the corpus-based models. This fact led us to propose the $Resnik_{ic-trebank-add1}$ and $Resnik_{ic-semcorraw-add1}$ IC models as baselines for any future study into the family of IC models and similarity measures.

6.3. Impact on the similarity measures

In Table 6, we can see that the $cosJ\&C$ similarity measure obtains the highest Pearson correlation results in the RG65 and SimLex665 datasets. In addition, this measure obtains the highest overall average Pearson and Spearman correlation values over all the datasets, followed by the FaITH, Meng and Gu [31] and Hadj Taieb et al. similarity measures. These results confirm and extend the conclusions in our previous aforementioned work.

In addition, the results in Table 5 and Fig. 3 show that the $cosJ\&C$ measure is the only IC-based similarity measure that obtains a statistically significant higher performance than the baseline defined by the J&C similarity measure. Therefore, the $cosJ\&C$ similarity measure obtains a statistically significant outperformance over the rest of state-of-the-art measures evaluated herein, and its overall performance also overcomes the Hadj Taieb et al. measure.

Despite the $cosJ\&C$ being a convincing winner according to the overall performance and the baseline defined by the J&C measure, it does not obtain statistically significant higher results when directly compared with the FaITH and Meng et al. [32] measures through a confidence interval analysis. We carried-out several confidence intervals to analyze changing the baseline that we do not show herein through lack of space. If the Lin measure is used as baseline, the FaITH and Meng et al. [32] similarity measures obtain a statistical significant higher performance than the Lin measure, but the $cosJ\&C$ does not. Using a pairwise similar confidence interval analysis to compare the performance of the FaITH, Meng et al. [32] and $cosJ\&C$ similarity measures we did not find statistically significant differences between them. In addition, in all the pairwise comparisons between the latter measures, the rest of the measures obtain a statistically significant lower performance than these three measures, except the Lin measure which does not exhibit a statistically significant difference as regard the $cosJ\&C$ measure.

In summary, the $cosJ\&C$ measure obtains the best overall performance, but it is not statistically different from the FaITH and Meng et al. [32] similarity measures, thus, these three measures are the winners in the family of IC-based similarity measures. However, the problem is still open. One advantage of the FaITH and Meng et al. measure over the $cosJ\&C$ measure is that the former ones exhibit a lower standard deviation (variance) as regard the IC models, as is shown in the last row of Table 5.

Looking at any result table in the appendix, we can appreciate that the Meng and Gu [31] and $cosJ\&C$ similarity measures always obtain the same Spearman correlation value as the Lin and Jiang-Conrath similarity measures for any IC model. This fact follows from that the Meng and Gu [31] and $cosJ\&C$ similarity measures are non-linear monotone transformations from the former ones, thus, the Spearman correlation does not change. The exponential-like transformations used by both measures contribute to improving the Pearson correlation value, however, the rank correlation remains invariant as was expected.

Looking at Table 5, we can appreciate that all the IC-based similarity measures exhibit strong performance dependence as regard the IC models. We can find the best IC model in bold within the column associated with each measure. On average, the best combinations of IC measures and models are as follows: (1) the Resnik measure and the Sánchez et al. [52] IC model, (2) the Lin measure and the Yuan et al. IC model, (3) the J&C measure and the Seco et al. IC model, (4) the P&S measure and the Seco et al. IC model, (5) the FaITH measure and the Yuan et al. IC model, (6) the Meng and Gu [31] measure and the Yuan et al. IC model, and (7) the $cosJ\&C$ measure and the Sánchez et al. [52] IC model.

In addition, the dependence on the IC models extends to the datasets, because the best performing IC model for each measure can change from one dataset to another. For instance, the $cosJ\&C$ obtains the highest Pearson correlation on the RG65 dataset with the Sánchez et al. IC model, however, on the SimLex665 dataset the best one is the Yuan et al. IC model. In the same way, the FaITH measure obtains its best results on the Agirre201, $P\&S_{full}$ and SimLex-665 datasets with the Yuan et al. IC model, however, on the MC28 and RG65 datasets it obtains its best results with the Sánchez et al. [52] and Zhou et al. IC models.

6.4. New state-of-the-art results

The Pearson and Spearman correlation values shown in Table 6 set the highest correlation values reported in the literature for the evaluation of the family of ontology-based similarity measures based on the same code implementation. The $cosJ\&C$ measure obtains the highest Pearson correlation values on the RG65 and SimLex665 dataset, whilst the Jiang-Conrath does the same on the MC28 dataset, the FaITH on the $P\&S_{full}$ dataset, and the Hadj Taieb et al. on the Agirre201 dataset. The $cosJ\&C$ similarity measure obtains Pearson and Spearman correlation values of 0.6106/0.6027 in the SimLex-665 (nouns) dataset. These results exceed the 0.599/0.591 values obtained by the best corpus-based method (UMBC introduced in Han et al. [16]) in a recent benchmark of corpus-based similarity measures [3, Table 1]. Therefore, the results obtained by the $cosJ\&C$ similarity measure are the highest Pearson and Spearman correlation values reported in the literature for any type of similarity measure on the noun database of SimLex-665.

6.5. Contradictory results

We confirm a contradictory result reported in Lastra-Díaz and García-Serrano [26]. In Meng and Gu [31], the authors report a Pearson correlation value of 0.8804 in the RG65 dataset for the intrinsic Seco et al. IC model, whilst we obtain 0.8596 herein, and other authors report 0.85 in [15, p. 256]. We subscribe to the warning about the reproducibility problems in the family of ontology-based similarity measures that is made in Fokkens et al. [11] and our aforementioned work. Thus, we invite to the research community to validate these contradictory results, as well as the rest of results reported herein.

Table 4

Best results for each IC model and dataset: Pearson (r) and Spearman (ρ) correlation coefficients, and averaged overall scores. The row are ordered according to the average Pearson correlation. Bold values represent the best score within each column.

Best results per dataset IC models	RG65		MC28		Agirre202		$P\&S_{full}$		SimLex665		Avg. overall scores		
	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ	both
Resnik _{ic-treebank-add1} [48]	0.8653	0.7831	0.8809	0.8882	0.6913	0.6461	0.9003	0.7783	0.5955	0.5810	0.7867	0.7353	0.7610
Yuan et al. [62]	0.8675	0.8206	0.8407	0.8274	0.7061	0.6656	0.9082	0.8199	0.6106	0.6027	0.7866	0.7473	0.7669
Seco et al. [56]	0.8642	0.8012	0.8557	0.8727	0.6969	0.6643	0.9042	0.7919	0.6048	0.5901	0.7852	0.7441	0.7646
Sánchez et al. [52]	0.8752	0.8034	0.8595	0.8492	0.6946	0.6576	0.9025	0.8003	0.5941	0.5906	0.7852	0.7402	0.7627
Meng et al. [32]	0.8723	0.8166	0.8393	0.8296	0.7039	0.6581	0.9057	0.8127	0.6010	0.5957	0.7844	0.7426	0.7635
Harispe (2012)	0.8589	0.7977	0.8575	0.8697	0.6960	0.6539	0.9003	0.7904	0.6056	0.5918	0.7836	0.7407	0.7622
Resnik _{ic-semcorraw-add1} [48]	0.8658	0.7922	0.8621	0.8712	0.6955	0.6505	0.8997	0.7835	0.5930	0.5782	0.7832	0.7351	0.7592
Sánchez et al. [51]	0.8616	0.7911	0.8507	0.8551	0.6973	0.6590	0.9042	0.7854	0.5995	0.5850	0.7827	0.7351	0.7589
CondProbCosine	0.8634	0.7896	0.8562	0.8606	0.6902	0.6524	0.9015	0.7834	0.5964	0.5828	0.7815	0.7337	0.7576
CondProbHypo	0.8658	0.8017	0.8552	0.8554	0.6874	0.6466	0.9015	0.7910	0.5940	0.5806	0.7808	0.7350	0.7579
CondProbLeaves	0.8635	0.7877	0.8511	0.8389	0.6891	0.6478	0.9008	0.7808	0.5934	0.5799	0.7796	0.7270	0.7533
CPCorpus _{ic-treebank-add1}	0.8633	0.7722	0.8678	0.8502	0.6807	0.6364	0.8987	0.7691	0.5863	0.5735	0.7794	0.7203	0.7498
CPCorpus _{ic-semcorraw-add1}	0.8647	0.7916	0.8504	0.8247	0.6792	0.6389	0.8979	0.7813	0.5843	0.5712	0.7753	0.7216	0.7484
Zhou et al. [63]	0.8589	0.8051	0.8403	0.8244	0.6848	0.6591	0.8905	0.7999	0.5985	0.5945	0.7746	0.7366	0.7556
CondProbLogistic _{k8}	0.8692	0.7993	0.8142	0.8034	0.6809	0.6460	0.9064	0.7921	0.5972	0.5791	0.7736	0.7240	0.7488
CondProbLogistic _{k10}	0.8689	0.7993	0.8109	0.8012	0.6784	0.6461	0.9067	0.7903	0.5964	0.5772	0.7722	0.7228	0.7475
CondProbLogistic _{k12}	0.8689	0.7948	0.8104	0.8023	0.6761	0.6444	0.9065	0.7915	0.5954	0.5763	0.7715	0.7219	0.7467
CondProbUniform	0.8425	0.7786	0.8039	0.7749	0.6516	0.6325	0.8644	0.7852	0.5416	0.5506	0.7408	0.7044	0.7226
Hadj Taieb et al. [15]	0.7933	0.7417	0.6899	0.6961	0.6490	0.6175	0.8167	0.7463	0.4921	0.4833	0.6882	0.6570	0.6726
Best values per dataset	0.8752	0.8206	0.8809	0.8882	0.7061	0.6656	0.9082	0.8199	0.6106	0.6027	0.7867	0.7473	0.7669

Table 5

Average on all the datasets of the Pearson (r) and Spearman (ρ) correlations for each pair (IC model, IC measure). Bold values represent the best score within each column.

Avg. values IC models	Resnik		Lin		J&C		P&S		FaITH (P&E)		Meng et al.		cosJ&C (L&G)		Average
	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ	
Seco (2004)	0.7385	0.6987	0.7731	0.7323	0.7705	0.7375	0.7784	0.7423	0.7743	0.7323	0.7758	0.7323	0.7790	0.7375	0.7699
Yuan (2013)	0.7372	0.6997	0.7744	0.7284	0.7613	0.7363	0.7745	.7402	0.7826	0.7284	0.7830	0.7284	0.7736	0.7363	0.7695
Meng (2012)	0.7367	0.6961	0.7704	0.7204	0.7632	0.7387	0.7708	0.7317	0.7794	0.7204	0.7790	0.7204	0.7758	0.7387	0.7679
Sánchez (2012)	0.7424	0.6957	0.7727	0.7264	0.7668	0.7281	0.7408	0.7148	0.7735	0.7264	0.7753	0.7264	0.7752	0.7281	0.7638
Resnik _{scr(1999)} scr(2008)	0.7449	0.7031	0.7672	0.7254	0.7669	0.7270	0.7402	0.7060	0.7707	0.7254	0.7718	0.7254	0.7732	0.7270	0.7621
Resnik _{trb(1999)} trb(2008)	0.7416	0.7005	0.7690	0.7219	0.7632	0.7303	0.7404	0.7008	0.7736	0.7219	0.7742	0.7219	0.7718	0.7303	0.7620
Sánchez (2011)	0.7459	0.7008	0.7612	0.7266	0.7681	0.7402	0.7174	0.6804	0.7780	0.7266	0.7752	0.7266	0.7811	0.7402	0.7610
Harispe (2012)	0.7302	0.6980	0.7735	0.7333	0.7599	0.7297	0.7455	0.7191	0.7711	0.7333	0.7739	0.7333	0.7674	0.7297	0.7602
CPCosine	0.7340	0.6923	0.7687	0.7229	0.7676	0.7300	0.7344	0.7124	0.7694	0.7229	0.7712	0.7229	0.7750	0.7300	0.7601
CPHypo	0.7330	0.6897	0.7673	0.7202	0.7659	0.7350	0.7332	0.7128	0.7688	0.7202	0.7705	0.7202	0.7739	0.7350	0.7589
CPProbLeaves	0.7326	0.6882	0.7670	0.7198	0.7641	0.7262	0.7314	0.7080	0.7691	0.7198	0.7705	0.7198	0.7726	0.7262	0.7582
CPCorpus _{scr}	0.7387	0.6971	0.7613	0.7116	0.7589	0.7179	0.7280	0.6950	0.7643	0.7116	0.7657	0.7116	0.7658	0.7179	0.7547
CPCorpus _{trb}	0.7353	0.6893	0.7630	0.7076	0.7553	0.7134	0.7288	0.6858	0.7675	0.7076	0.7681	0.7076	0.7641	0.7134	0.7546
Zhou (2008)	0.7236	0.6974	0.7418	0.7250	0.7411	0.7353	0.7497	0.7267	0.7734	0.7250	0.7676	0.7250	0.7645	0.7353	0.7517
CPLog _{k8}	0.7142	0.6938	0.7659	0.7178	0.7296	0.7189	0.7125	0.6888	0.7714	0.7178	0.7727	0.7178	0.7477	0.7189	0.7449
CPLog _{k10}	0.7098	0.6882	0.7657	0.7137	0.7284	0.7178	0.7125	0.6877	0.7698	0.7136	0.7715	0.7137	0.7461	0.7178	0.7434
CPLog _{k12}	0.7059	0.6866	0.7651	0.7148	0.7265	0.7149	0.7116	0.6917	0.7684	0.7148	0.7704	0.7148	0.7440	0.7149	0.7417
CPProbUnif	0.6135	0.6166	0.7031	0.6987	0.6376	0.6818	0.6037	0.6063	0.7362	0.6987	0.7297	0.6987	0.6557	0.6817	0.6685
Hadj Taieb (2014)	0.4233	0.5364	0.6765	0.6570	0.3267	0.4627	0.4196	0.5164	0.6882	0.6570	0.6857	0.6570	0.3278	0.4627	0.5068
Best values	0.7459	0.7031	0.7744	0.7333	0.7705	0.7402	0.7784	0.7423	0.7826	0.7333	0.7830	0.7333	0.7811	0.7402	
Standard dev	0.0753	0.0400	0.0255	0.0168	0.1018	0.0618	0.0802	0.0520	0.0210	0.0169	0.0221	0.0168	0.1033	0.0618	

Table 6

Best results for each measure and dataset: Pearson (r) and Spearman (ρ) correlation coefficients, and averaged overall scores. Bold values represent the best score within each column.

Best results IC-based measures	RG65		MC28		Agirre202		$P\&S_{full}$		SimLex665		Avg. overall scores		
	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ	both
cosJ&C [26]	0.8752	0.8166	0.8710	0.8882	0.6904	0.6612	0.8996	0.8127	0.6106	0.6027	0.7893	0.7563	0.7728
FaITH [44]	0.8683	0.7977	0.8344	0.8403	0.7061	0.6652	0.9082	0.7936	0.6063	0.5918	0.7847	0.7377	0.7612
Meng-Gu [31]	0.8692	0.7977	0.8361	0.8403	0.7040	0.6652	0.9064	0.7936	0.6062	0.5918	0.7844	0.7377	0.7611
Hadj Taieb et al. [15]	0.8670	0.7972	0.8248	0.8077	0.7123	0.6633	0.9068	0.7973	0.6093	0.5960	0.7840	0.7323	0.7582
Jiang-Conrath (1997)	0.8619	0.8166	0.8809	0.8882	0.6724	0.6612	0.8825	0.8127	0.6027	0.6027	0.7801	0.7563	0.7682
Lin [30]	0.8689	0.7977	0.8393	0.8403	0.6870	0.6652	0.8953	0.7936	0.6045	0.5918	0.7790	0.7377	0.7584
Pirró-Seco [45]	0.8622	0.8206	0.8466	0.8678	0.6890	0.6656	0.8970	0.8199	0.5981	0.5869	0.7786	0.7522	0.7654
Resnik [47]	0.8409	0.7833	0.8202	0.8296	0.6760	0.6481	0.8829	0.7800	0.5506	0.5346	0.7541	0.7151	0.7346
Best values per dataset	0.8752	0.8206	0.8809	0.8882	0.7123	0.6656	0.9082	0.8199	0.6106	0.6027	0.7893	0.7563	0.7728

Table 7 Pearson (r) and Spearman (ρ) correlation coefficients for all the IC models and measures in the RG65 dataset using Wordnet 3.0. Bold values represent the best overall scores in dataset.

RG65 dataset	Resnik		Lin		J&C		P&S		FaITH (P&E)		Meng et al.		cosJ&C (L&G)		Best row values	
	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ
Seco (2004)	0.8326	0.7735	0.8609	0.7972	0.8546	0.7866	0.8622	0.8012	0.8565	0.7972	0.8596	0.7972	0.8642	0.8012	0.8642	0.8012
Zhou (2008)	0.8080	0.7690	0.8259	0.7881	0.8286	0.8051	0.8334	0.7958	0.8589	0.7881	0.8539	0.7881	0.8558	0.8051	0.8589	0.8051
Sánchez (2011)	0.8409	0.7714	0.8530	0.7944	0.8619	0.8034	0.8105	0.7671	0.8683	0.7944	0.8663	0.7944	0.8752	0.8034	0.8752	0.8034
Hanispe (2012)	0.8203	0.7601	0.8589	0.7977	0.8460	0.7883	0.8348	0.7866	0.8509	0.7977	0.8553	0.7977	0.8546	0.7883	0.8589	0.7977
Sánchez (2012)	0.8355	0.7706	0.8616	0.7911	0.8508	0.7779	0.8332	0.7856	0.8565	0.7911	0.8600	0.7911	0.8606	0.7779	0.8616	0.7911
Meng (2012)	0.8260	0.7607	0.8608	0.7817	0.8598	0.8166	0.8586	0.8140	0.8658	0.7817	0.8670	0.7817	0.8723	0.8166	0.8723	0.8166
Yuan (2013)	0.8243	0.7742	0.8621	0.7919	0.8505	0.8050	0.8607	0.8206	0.8649	0.7919	0.8675	0.7919	0.8632	0.8050	0.8675	0.8206
Hadij Taieb (2014)	0.4658	0.5990	0.7825	0.7417	0.4543	0.6126	0.5090	0.6123	0.7933	0.7417	0.7924	0.7417	0.4552	0.6126	0.7933	0.7417
Resnik _{scr(1999)}	0.8363	0.7788	0.8569	0.7922	0.8595	0.7810	0.8342	0.7820	0.8554	0.7922	0.8581	0.7922	0.8658	0.7810	0.8658	0.7810
Resnik _{scr(2008)}	0.8345	0.7777	0.8589	0.7761	0.8561	0.7831	0.8335	0.7633	0.8585	0.7761	0.8609	0.7761	0.8653	0.7831	0.8653	0.7831
Resnik _{trb(2008)}	0.8345	0.7777	0.8589	0.7761	0.8561	0.7831	0.8335	0.7633	0.8585	0.7761	0.8609	0.7761	0.8653	0.7831	0.8653	0.7831
IC models introduced in this work. (scr = ic-semconrow-add1, trb = ic-treebank-add1)																
New IC models	0.8283	0.7561	0.8587	0.7888	0.8562	0.8017	0.8286	0.7941	0.8556	0.7888	0.8585	0.7888	0.8658	0.8017	0.8658	0.8017
CondProbHypo	0.6844	0.7003	0.7958	0.7786	0.7228	0.7613	0.6840	0.6930	0.8425	0.7786	0.8323	0.7786	0.7473	0.7613	0.8425	0.7786
CondProbUnif	0.8272	0.7549	0.8575	0.7868	0.8535	0.7877	0.8263	0.7848	0.8550	0.7868	0.8578	0.7868	0.8635	0.7877	0.8635	0.7877
CondProbLeaves	0.8021	0.7833	0.8681	0.7966	0.8260	0.7993	0.8147	0.7792	0.8652	0.7966	0.8692	0.7966	0.8501	0.7993	0.8692	0.7993
CondProbLog ₈	0.7998	0.7778	0.8689	0.7934	0.8236	0.7993	0.8138	0.7782	0.8645	0.7934	0.8689	0.7934	0.8478	0.7993	0.8689	0.7993
CondProbLog ₁₀	0.7972	0.7773	0.8689	0.7948	0.8206	0.7942	0.8119	0.7810	0.8637	0.7948	0.8684	0.7948	0.8452	0.7942	0.8689	0.7948
CondProbLog ₁₂	0.8294	0.7710	0.8591	0.7896	0.8265	0.7835	0.8265	0.7830	0.8549	0.7896	0.8581	0.7896	0.8634	0.7835	0.8634	0.7896
CondProbCosine	0.8321	0.7760	0.8549	0.7806	0.8568	0.7916	0.8250	0.7763	0.8549	0.7806	0.8571	0.7806	0.8647	0.7916	0.8647	0.7916
CPCorpus _{scr(2008)}	0.8296	0.7702	0.8560	0.7652	0.8527	0.7722	0.8223	0.7515	0.8575	0.7652	0.8591	0.7652	0.8633	0.7722	0.8633	0.7722
CPCorpus _{trb(2008)}	0.8409	0.7833	0.8689	0.7977	0.8619	0.8166	0.8622	0.8206	0.8683	0.7977	0.8692	0.7977	0.8752	0.8166	0.8752	0.8206

Table 8 Pearson (r) and Spearman (ρ) correlation coefficients for all the IC models and measures in the MC28 dataset using Wordnet 3.0. Bold values represent the best overall scores in dataset.

MC28 dataset	Resnik		Lin		J&C		P&S		FaITH (P&E)		Meng et al.		cosJ&C (L&G)		Best row values	
	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ
Seco (2004)	0.7834	0.7947	0.8240	0.8314	0.8557	0.8727	0.8463	0.8678	0.8094	0.8314	0.8144	0.8314	0.8427	0.8277	0.8557	0.8727
Zhou (2008)	0.8179	0.7971	0.8222	0.8072	0.8282	0.8244	0.8281	0.8091	0.8344	0.8072	0.8338	0.8072	0.8403	0.8244	0.8403	0.8244
Sánchez (2011)	0.8189	0.7937	0.8357	0.8114	0.8595	0.8492	0.8056	0.7348	0.8343	0.8114	0.8359	0.8114	0.8476	0.8492	0.8595	0.8492
Hanispe (2012)	0.7705	0.8280	0.8236	0.8328	0.8575	0.8697	0.8299	0.8339	0.8064	0.8328	0.8122	0.8328	0.8497	0.8697	0.8575	0.8697
Sánchez (2012)	0.7905	0.7774	0.8261	0.8212	0.8507	0.8551	0.8177	0.8329	0.8103	0.8212	0.8159	0.8212	0.8411	0.8551	0.8507	0.8551
Meng (2012)	0.8202	0.8296	0.8393	0.8314	0.8314	0.8198	0.8330	0.8050	0.8330	0.8080	0.8361	0.8080	0.8393	0.8198	0.8393	0.8296
Yuan (2013)	0.8084	0.7971	0.8341	0.8042	0.8347	0.8274	0.8384	0.8083	0.8277	0.8042	0.8315	0.8042	0.8407	0.8274	0.8407	0.8274
Hadij Taieb (2014)	0.5391	0.6340	0.6842	0.6961	0.4587	0.6102	0.4845	0.4673	0.6899	0.6961	0.6875	0.6961	0.4594	0.6102	0.6899	0.6961
Resnik _{scr(1999)}	0.7894	0.7905	0.8245	0.8225	0.8621	0.8712	0.8267	0.8108	0.8101	0.8225	0.8155	0.8225	0.8498	0.8712	0.8621	0.8712
Resnik _{scr(2008)}	0.7929	0.7935	0.8350	0.8403	0.8809	0.8882	0.8466	0.8255	0.8233	0.8403	0.8276	0.8403	0.8710	0.8882	0.8809	0.8882
Resnik _{trb(2008)}	0.7929	0.7935	0.8350	0.8403	0.8809	0.8882	0.8466	0.8255	0.8233	0.8403	0.8276	0.8403	0.8710	0.8882	0.8809	0.8882
IC models introduced in this work. (scr = ic-semconrow-add1, trb = ic-treebank-add1)																
New IC models	0.7860	0.8131	0.8215	0.8034	0.8552	0.8554	0.8110	0.8187	0.8070	0.8034	0.8124	0.8034	0.8429	0.8554	0.8552	0.8554
CondProbHypo	0.7408	0.7564	0.7753	0.7749	0.6483	0.7281	0.6365	0.6155	0.8039	0.7749	0.7971	0.7749	0.6698	0.7281	0.8039	0.7749
CondProbUnif	0.7869	0.8073	0.8219	0.8039	0.8511	0.8389	0.8073	0.8110	0.8080	0.8039	0.8131	0.8039	0.8402	0.8389	0.8511	0.8389
CondProbLeaves	0.7700	0.7909	0.8142	0.7752	0.8020	0.8034	0.7687	0.7377	0.8079	0.7752	0.8119	0.7752	0.8095	0.8034	0.8142	0.8034
CondProbLog ₈	0.7579	0.7659	0.8109	0.7612	0.8036	0.8012	0.7685	0.7344	0.8041	0.7612	0.8083	0.7612	0.8105	0.8012	0.8109	0.8012
CondProbLog ₁₀	0.7488	0.7659	0.8082	0.7667	0.8035	0.8023	0.7673	0.7470	0.8013	0.7667	0.8056	0.7667	0.8104	0.8023	0.8104	0.8023
CondProbLog ₁₂	0.7797	0.7791	0.8208	0.8116	0.8562	0.8606	0.8127	0.8324	0.8054	0.8116	0.8109	0.8116	0.8445	0.8606	0.8606	0.8606
CondProbCosine	0.7890	0.7849	0.8202	0.7943	0.8504	0.8247	0.8086	0.7889	0.8062	0.7943	0.8118	0.7943	0.8422	0.8247	0.8504	0.8247
CPCorpus _{scr(2008)}	0.7912	0.7707	0.8290	0.8058	0.8678	0.8502	0.8281	0.7946	0.8163	0.8058	0.8209	0.8058	0.8613	0.8502	0.8678	0.8502
CPCorpus _{trb(2008)}	0.8202	0.8296	0.8393	0.8403	0.8809	0.8882	0.8466	0.8255	0.8233	0.8403	0.8276	0.8403	0.8710	0.8882	0.8809	0.8882
Best by measure	0.8202	0.8296	0.8393	0.8403	0.8809	0.8882	0.8466	0.8255	0.8233	0.8403	0.8276	0.8403	0.8710	0.8882	0.8809	0.8882

Table 9 Pearson (r) and Spearman (ρ) correlation coefficients for all the IC models and measures in the Agirre201 dataset using Wordnet 3.0. Bold values represent the best overall scores in dataset.

Agirre201 dataset	Resnik		Lin		J&C		P&S		FaITH (P&E)		Meng et al.		cosj&c (L&G)		Best row values	
	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ
Seco (2004)	0.6629	0.6308	0.6850	0.6530	0.6724	0.6612	0.6890	0.6643	0.6966	0.6530	0.6969	0.6530	0.6904	0.6612	0.6969	0.6643
Zhou (2008)	0.6453	0.6460	0.6409	0.6591	0.6288	0.6524	0.6503	0.6581	0.6848	0.6591	0.6753	0.6591	0.6564	0.6524	0.6848	0.6591
Sánchez (2011)	0.6592	0.6408	0.6643	0.6534	0.6591	0.6576	0.6339	0.6224	0.6946	0.6534	0.6889	0.6534	0.6890	0.6576	0.6946	0.6576
Harispe (2012)	0.6591	0.6309	0.6870	0.6539	0.6501	0.6429	0.6581	0.6454	0.6941	0.6539	0.6960	0.6539	0.6647	0.6429	0.6960	0.6539
Sánchez (2012)	0.6678	0.6345	0.6848	0.6494	0.6720	0.6590	0.6602	0.6399	0.6972	0.6494	0.6973	0.6494	0.6878	0.6590	0.6973	0.6590
Meng (2012)	0.6756	0.6462	0.6817	0.6581	0.6593	0.6488	0.6863	0.6560	0.7039	0.6581	0.7008	0.6581	0.6747	0.6488	0.7039	0.6581
Yuan (2013)	0.6760	0.6481	0.6858	0.6652	0.6543	0.6505	0.6863	0.6656	0.7061	0.6652	0.7040	0.6652	0.6695	0.6505	0.7061	0.6656
Hadji Taieb (2014)	0.4165	0.5343	0.6345	0.6175	0.0790	0.1556	0.3150	0.4676	0.6490	0.6175	0.6467	0.6175	0.6793	0.1556	0.6490	0.6175
Resnik _{K₁(1999)}	0.6688	0.6314	0.6783	0.6505	0.6538	0.6437	0.6526	0.6359	0.6955	0.6505	0.6943	0.6505	0.6708	0.6437	0.6955	0.6505
Resnik _{K₁(2008)}	0.6716	0.6320	0.6697	0.6461	0.6349	0.6382	0.6401	0.6282	0.6913	0.6461	0.6882	0.6461	0.6524	0.6382	0.6913	0.6461
IC models introduced in this work. (scr = ic-semcorraw-add1, trb = ic-treebank-add1)																
New IC models	0.6557	0.6267	0.6751	0.6466	0.6585	0.6462	0.6470	0.6329	0.6868	0.6466	0.6874	0.6466	0.6748	0.6462	0.6874	0.6466
CondProbHypo	0.5713	0.5617	0.6156	0.6325	0.5597	0.6005	0.5682	0.5704	0.6516	0.6325	0.6432	0.6325	0.5719	0.6002	0.6516	0.6325
CondProbUnif	0.6569	0.6289	0.6765	0.6478	0.6601	0.6476	0.6478	0.6363	0.6888	0.6478	0.6891	0.6478	0.6764	0.6476	0.6891	0.6478
CondProbLeaves	0.6501	0.6241	0.6618	0.6460	0.6206	0.6302	0.6290	0.6284	0.6809	0.6460	0.6796	0.6460	0.6380	0.6302	0.6809	0.6460
CondProbLog _{k=8}	0.6453	0.6219	0.6615	0.6461	0.6215	0.6304	0.6304	0.6280	0.6784	0.6461	0.6778	0.6461	0.6378	0.6304	0.6784	0.6461
CondProbLog _{k=10}	0.6411	0.6165	0.6606	0.6444	0.6212	0.6296	0.6311	0.6286	0.6761	0.6444	0.6761	0.6444	0.6369	0.6296	0.6761	0.6444
CondProbLog _{k=12}	0.6579	0.6274	0.6789	0.6479	0.6660	0.6524	0.6521	0.6388	0.6893	0.6479	0.6902	0.6479	0.6805	0.6524	0.6902	0.6524
CondProbCosine	0.6603	0.6285	0.6656	0.6389	0.6419	0.6279	0.6377	0.6214	0.6788	0.6389	0.6792	0.6389	0.6555	0.6279	0.6792	0.6389
CPCorpus _{scr(2008)}	0.6638	0.6282	0.6603	0.6364	0.6285	0.6256	0.6326	0.6202	0.6807	0.6364	0.6779	0.6364	0.6426	0.6256	0.6807	0.6364
CPCorpus _{trb(2008)}	0.6760	0.6481	0.6870	0.6652	0.6724	0.6612	0.6890	0.6656	0.7061	0.6652	0.7040	0.6652	0.6904	0.6612	0.7061	0.6656

7. Conclusions and future work

We have introduced five new intrinsic IC models and one new corpus-based IC model based on the preservation of the probabilistic structure, and the integration of a notion of cognitive similarity inspired by cognitive evidence. The proposed approach defines an open framework for the development of new intrinsic IC models based on alternative forms of estimating the conditional probabilities between concepts.

Most of new intrinsic IC models rival the state-of-the-art models, with the exception of the naive *CondProbUniform* model. The integration of the probabilistic structure in the IC models has proven to be helpful in getting results rivaling the state-of-the-art IC models, but it has not been enough to exceed the state of the art by itself. Nevertheless, we expect that the encoding of the structure axioms into the IC models contributes to a better understanding of the problem, as well as the start of a line of research in conditional probability estimation. On the other hand, the results of the *CondProbCosine* and *CondProbLogistic* model confirm that the encoding of cognitive similarity notions within the IC models and measures is a line of research that deserves to be explored.

We have proved that most intrinsic IC models and IC-based similarity measures do not exhibit significant statistical differences as regard the baselines of the experiments. Despite the Seco et al. IC model obtaining the highest overall average correlation values, the statistical evidence proves that the Sánchez et al. [53] IC model obtains a significant statistical outperformance over the baseline and the rest of the IC models, this latter model being the IC model that best generalizes any IC-based similarity measure. We prove that the cosj&c similarity measure obtains the best overall results, obtaining a significant statistical outperformance over the rest of the IC models and measures in comparison with the baseline. However, a deeper confidence interval analysis between the FaITH, Meng and Gu [31] and cosj&c similarity measures confirm that there is no a statistically significant difference between them.

The lack of a statistically significant difference between most intrinsic IC models and the corpus-based IC model *Resnik_{ic-treebank-add1}* defined as the baseline allows the following conclusions to be extracted: (1) this fact refutes a previous belief about the outperformance of the intrinsic IC models over the corpus-based ones, confirming the same finding in our aforementioned work, and (2) this fact confirms the achievements of the family of intrinsic IC models, which offers a practical alternative to the corpus-based models without a significant reduction in performance. Among the set of rivaling state-of-the-art intrinsic IC models we have the Seco et al., Yuan et al., Meng et al., Sánchez et al. [52], Harispe, *CondProbCosine*, *CondProbHypo*, and Zhou IC models. The statistical significance of the results confirms that most of the IC models offer similar results, and the problem is still open.

As forthcoming activities, we would like to carry-out an in-depth study into the relationship between the corpus-based IC models evaluated herein and the intrinsic IC models. We plan to study the fitting quality of the intrinsic IC models based on a direct comparison between IC models using a well-defined metric on function spaces, or some correlation measure, such as that proposed in Blanchard et al. [5]. In addition, we would also like to evaluate the intrinsic IC models introduced in this latter work.

Acknowledgements

Despite deciding to develop our own software library to implement all the IC-based models and measures evaluated in this work, we would like to express our gratitude to Sébastien Harispe, who provided us the source code of the SML library, offering his total

Table 10 Pearson (r) and Spearman (ρ) correlation coefficients for all the IC models and measures in the $P_{\mathcal{L}}S_{full}$ dataset using Wordnet 3.0. Bold values represent the best overall scores in dataset.

$P_{\mathcal{L}}S_{full}$ dataset	Resnik		Lin		J&C		P&S		FaITH (P&E)		Meng et al.		cosJ&C (L&G)		Best row values	
	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ
Seco (2004)	0.8799	0.7735	0.8945	0.7911	0.8781	0.7768	0.8970	0.7919	0.9042	0.7911	0.9031	0.7911	0.8966	0.7768	0.9042	0.7919
Zhou (2008)	0.8357	0.7720	0.8420	0.7867	0.8372	0.7999	0.8563	0.7938	0.8905	0.7867	0.8806	0.7867	0.8726	0.7999	0.8905	0.7999
Sánchez (2011)	0.8740	0.7732	0.8738	0.7936	0.8762	0.8003	0.8051	0.7573	0.9025	0.7936	0.8964	0.7936	0.8996	0.8003	0.9025	0.8003
Hanispe (2012)	0.8699	0.7517	0.8933	0.7904	0.8686	0.7741	0.8411	0.7680	0.9001	0.7904	0.9003	0.7904	0.8834	0.7741	0.9003	0.7904
Sánchez (2012)	0.8829	0.7727	0.8948	0.7854	0.8742	0.7652	0.8398	0.7681	0.9042	0.7854	0.9035	0.7854	0.8918	0.7652	0.9042	0.7854
Meng (2012)	0.8645	0.7543	0.8863	0.7776	0.8715	0.8127	0.8897	0.8122	0.9057	0.7776	0.9025	0.7776	0.8917	0.8127	0.9057	0.8127
Yuan (2013)	0.8655	0.7759	0.8896	0.7905	0.8641	0.7957	0.8891	0.8199	0.9082	0.7905	0.9061	0.7905	0.8840	0.7957	0.9082	0.8199
Hadij Taieb (2014)	0.4915	0.6140	0.7962	0.7463	0.4261	0.6094	0.4660	0.6107	0.8167	0.7463	0.8125	0.7463	0.4272	0.6094	0.8167	0.7463
Resnik _{scr(1999)}	0.8972	0.7800	0.8876	0.7835	0.8825	0.7653	0.8370	0.7604	0.8997	0.7835	0.8983	0.7835	0.8968	0.7653	0.8997	0.7835
Resnik _{scr(1999)}	0.8736	0.7783	0.8877	0.7660	0.8749	0.7717	0.8328	0.7422	0.9003	0.7660	0.8988	0.7660	0.8922	0.7717	0.9003	0.7783
Resnik _{trh(2008)}	0.8736	0.7783	0.8877	0.7660	0.8749	0.7717	0.8328	0.7422	0.9003	0.7660	0.8988	0.7660	0.8922	0.7717	0.9003	0.7783
IC models introduced in this work. (scr = ic-semcrraw-add1, trh = ic-treebank-add1)																
New IC models	0.8725	0.7461	0.8903	0.7824	0.8783	0.7910	0.8325	0.7790	0.9015	0.7824	0.9002	0.7824	0.8963	0.7910	0.9015	0.7910
CondProbHypo	0.7077	0.6997	0.8042	0.7852	0.7204	0.7684	0.6602	0.6905	0.8644	0.7852	0.8498	0.7852	0.7478	0.7684	0.8644	0.7852
CondProbUniform	0.8715	0.7457	0.8887	0.7808	0.8748	0.7766	0.8297	0.7689	0.9008	0.7808	0.8992	0.7808	0.8930	0.7766	0.9008	0.7808
CondProbLeaves	0.8419	0.7783	0.8928	0.7921	0.8315	0.7880	0.8061	0.7630	0.9064	0.7921	0.9056	0.7921	0.8632	0.7880	0.9064	0.7921
CondProbLog ₈	0.8398	0.7793	0.8946	0.7903	0.8302	0.7884	0.8069	0.7622	0.9067	0.7903	0.9063	0.7903	0.8614	0.7884	0.9067	0.7903
CondProbLog ₁₀	0.8373	0.7789	0.8953	0.7915	0.8278	0.7831	0.8061	0.7657	0.9065	0.7915	0.9064	0.7915	0.8591	0.7831	0.9065	0.7915
CondProbLog ₁₂	0.8717	0.7689	0.8915	0.7834	0.8778	0.7710	0.8318	0.7664	0.9015	0.7834	0.9005	0.7834	0.8943	0.7710	0.9015	0.7834
CondProbCosine	0.8751	0.7740	0.8843	0.7730	0.8779	0.7813	0.8253	0.7569	0.8979	0.7730	0.8961	0.7730	0.8937	0.7813	0.8979	0.7813
CPCorpus _{scr(2008)}	0.8649	0.7691	0.8840	0.7572	0.8711	0.7615	0.8203	0.7291	0.8987	0.7572	0.8964	0.7572	0.8892	0.7615	0.8987	0.7691
CPCorpus _{trh(2008)}	0.8829	0.7800	0.8953	0.7936	0.8825	0.8127	0.8970	0.8199	0.9082	0.7936	0.9064	0.7936	0.8996	0.8127	0.9082	0.8199

Table 11 Pearson (r) and Spearman (ρ) correlation coefficients for all the IC models and measures in the SimLex-665 dataset using Wordnet 3.0. Bold values represent the best overall scores in dataset.

SimLex665	Resnik		Lin		J&C		P&S		FaITH (P&E)		Meng et al.		cosJ&C (L&G)		Best row values	
	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ
Seco (2004)	0.5339	0.5210	0.6010	0.5888	0.5918	0.5901	0.5975	0.5862	0.6046	0.5888	0.6048	0.5888	0.6013	0.5901	0.6048	0.5901
Zhou (2008)	0.5110	0.5026	0.5778	0.5841	0.5825	0.5945	0.5803	0.5767	0.5985	0.5841	0.5945	0.5841	0.5973	0.5945	0.5985	0.5945
Sánchez (2011)	0.5364	0.5248	0.5791	0.5803	0.5838	0.5906	0.5318	0.5204	0.5904	0.5803	0.5883	0.5803	0.5941	0.5906	0.5941	0.5906
Hanispe (2012)	0.5310	0.5192	0.6045	0.5918	0.5775	0.5735	0.5634	0.5617	0.6039	0.5918	0.6056	0.5918	0.5844	0.5735	0.6056	0.5918
Sánchez (2012)	0.5354	0.5232	0.5964	0.5850	0.5861	0.5835	0.5532	0.5476	0.5991	0.5850	0.5995	0.5850	0.5945	0.5835	0.5995	0.5850
Meng (2012)	0.4972	0.4896	0.5841	0.5767	0.5939	0.5957	0.5863	0.5714	0.5887	0.5767	0.5884	0.5767	0.6010	0.5957	0.6010	0.5957
Yuan (2013)	0.5120	0.5030	0.6004	0.5903	0.6027	0.6027	0.5981	0.5869	0.6063	0.5903	0.6062	0.5903	0.6106	0.6027	0.6106	0.6027
Hadij Taieb (2014)	0.2033	0.3004	0.4849	0.4833	0.2154	0.3256	0.3237	0.4243	0.4921	0.4833	0.4896	0.4833	0.2177	0.3256	0.4921	0.4833
Resnik _{scr(1999)}	0.5506	0.5346	0.5888	0.5782	0.5768	0.5736	0.5507	0.5408	0.5930	0.5782	0.5926	0.5782	0.5832	0.5736	0.5930	0.5782
Resnik _{scr(1999)}	0.5355	0.5210	0.5935	0.5810	0.5692	0.5700	0.5489	0.5450	0.5948	0.5810	0.5955	0.5810	0.5781	0.5700	0.5955	0.5810
Resnik _{trh(2008)}	0.5355	0.5210	0.5935	0.5810	0.5692	0.5700	0.5489	0.5450	0.5948	0.5810	0.5955	0.5810	0.5781	0.5700	0.5955	0.5810
IC models introduced in this work. (scr = ic-semcrraw-add1, trh = ic-treebank-add1)																
New IC models	0.5223	0.5067	0.5910	0.5799	0.5811	0.5806	0.5468	0.5391	0.5932	0.5798	0.5940	0.5799	0.5896	0.5806	0.5940	0.5806
CondProbHypo	0.5223	0.5067	0.5910	0.5799	0.5811	0.5806	0.5468	0.5391	0.5932	0.5798	0.5940	0.5799	0.5896	0.5806	0.5940	0.5806
CondProbUniform	0.5204	0.3649	0.5245	0.5223	0.5366	0.5506	0.4693	0.4620	0.5186	0.5223	0.5260	0.5223	0.5416	0.5506	0.5416	0.5506
CondProbLeaves	0.5068	0.4925	0.5926	0.5791	0.5808	0.5799	0.5457	0.5392	0.5797	0.5796	0.5934	0.5796	0.5897	0.5799	0.5934	0.5799
CondProbLog ₈	0.5065	0.4961	0.5927	0.5772	0.5679	0.5738	0.5441	0.5358	0.5966	0.5790	0.5972	0.5791	0.5778	0.5738	0.5972	0.5791
CondProbLog ₁₀	0.5054	0.4945	0.5923	0.5763	0.5691	0.5653	0.5431	0.5358	0.5955	0.5771	0.5964	0.5772	0.5727	0.5696	0.5964	0.5772
CondProbLog ₁₂	0.5074	0.4945	0.5923	0.5763	0.5691	0.5653	0.5416	0.5361	0.5942	0.5763	0.5954	0.5763	0.5684	0.5653	0.5954	0.5763
CondProbCosine	0.5404	0.5221	0.5816	0.5712	0.5675	0.5640	0.5434	0.5317	0.5837	0.5712	0.5843	0.5712	0.5731	0.5640	0.5843	0.5712
CPCorpus _{scr(2008)}	0.5271	0.5085	0.5856	0.5735	0.5564	0.5578	0.5404	0.5335	0.5844	0.5735	0.5863	0.5735	0.5642	0.5578	0.5863	0.5735
CPCorpus _{trh(2008)}	0.5271	0.5085	0.5856	0.5735	0.5564	0.5578	0.5404	0.5335	0.5844	0.5735	0.5863	0.5735	0.5642	0.5578	0.5863	0.5735
Best by measure	0.5506	0.5346	0.6045	0.5918	0.6027	0.6027	0.5981	0.5869	0.6063	0.5918	0.6062	0.5918	0.6106	0.6027	0.6106	0.6027

support. Mohamed Hadj Taieb kindly offered us his total support to replicate their similarity measure exactly. Ted Pedersen kindly answered all our questions and provided us with the WordNet-based frequency files used to build all the corpus-based IC models used in our experiments. Giuseppe Pirró and Rajendra Banjade kindly answered all our questions to clarify the corpus-based IC models used in their experiments. Alexis Moreno-Pulido helped us to find some old papers. Mark Hallett reviewed the English translation. Finally, we would like to thank the anonymous reviewers whose comments have improved the quality of the paper. To all of them, we would like to express our most sincere gratitude. This work has been partially supported by the Spanish VOXPOPULI(TIN2013-47090-C3-1-P) Project.

Appendix: experimental results

Tables 4–11.

References

- [1] E. Agirre, E. Alfonseca, K. Hall, J. Kravalova, M. Paşca, A. Soroa, A study on similarity and relatedness using distributional and WordNet-based approaches, in: *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL '09)*, Stroudsburg, PA, USA, 2009, pp. 19–27.
- [2] R.B. Ash, C.A. Doléans-Dade, *Probability & Measure Theory*, second ed., Academic Press, 2000.
- [3] R. Banjade, N. Maharjan, N.B. Niraula, V. Rus, D. Gautam, Lemon and tea are not similar: measuring word-to-word similarity by combining different methods, in: A. Gelbukh (Ed.), *Proc. of the 16th CICLing International Conference on Computational Linguistics and Intelligent Text Processing, LNCS*, vol. 9041, Springer, Cairo, Egypt, 2015, pp. 335–346.
- [4] M. Batet, S. Harispe, S. Ranwez, D. Sánchez, V. Ranwez, An information theoretic approach to improve semantic similarity assessments across multiple ontologies, *Inform. Sci.* 283 (0) (2014) 197–210.
- [5] E. Blanchard, M. Harzallah, P. Kuntz, A generic framework for comparing semantic similarities on a subsumption hierarchy, in: M. Ghallab, C.D. Spyropoulos, N. Fakotakis, N. Avouris (Eds.), *Proceedings of the ECAI Frontiers in Artificial Intelligence and Applications*, vol. 178, IOS Press, 2008, pp. 20–24.
- [6] A. Budanitsky, G. Hirst, Evaluating WordNet-based measures of lexical semantic relatedness, *Comput. Linguist.* 32 (1) (2006) 13–47.
- [7] F.M. Couto, H.S. Pinto, The next generation of similarity measures that fully explore the semantics in biomedical ontologies, *J. Bioinform. Comput. Biol.* 11 (5) (2013) 1371001.
- [8] V. Cross, X. Hu, Using semantic similarity in ontology alignment, in: *Proc. of the Sixth International Workshop on Ontology Matching (OM)*, 10th Intl. Semantic Web Conference (ISWC 2011), Bonn, Germany, 2011, pp. 61–72.
- [9] V. Cross, X. Yu, X. Hu, Unifying ontological similarity measures: a theoretical and empirical investigation, *Int. J. Approx. Reason.* 54 (7) (2013) 861–875.
- [10] J.B. du Prel, G. Hommel, B. Röhrig, M. Blettner, Confidence interval or p-value?: part 4 of a series on evaluation of scientific publications, *Deutsches Arzteblatt Int.* 106 (19) (2009) 335–339.
- [11] A. Fokkens, M. Van Erp, M. Postma, T. Pedersen, P. Vossen, N. Freire, Offspring from reproduction problems: what replication failure teaches us, in: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL*, Sofia, Bulgaria, 2013, pp. 1691–1701.
- [12] J.B. Gao, B.W. Zhang, X.H. Chen, A WordNet-based semantic similarity measurement combining edge-counting and information content theory, *Eng. Appl. Artif. Intell.* 39 (0) (2015) 80–88.
- [13] P. Gärdenfors, *The Geometry of Meaning: Semantics Based on Conceptual Spaces*, MIT Press, 2014.
- [14] M.A. Hadj Taieb, M. Ben Aouicha, A. Ben Hamadou, A new semantic relatedness measurement using WordNet features, *Knowl. Inform. Syst.* 41 (2) (2014) 467–497.
- [15] M.A. Hadj Taieb, M. Ben Aouicha, A. Ben Hamadou, Ontology-based approach for measuring semantic similarity, *Eng. Appl. Artif. Intell.* 36 (0) (2014) 238–261.
- [16] L. Han, A. Kashyap, T. Finin, J. Mayfield, J. Weese, UMBC EBILITY-CORE: semantic textual similarity systems, *Proceedings of the Second Joint Conference on Lexical and Computational Semantics*, vol. 1, ACL, Atlanta, Georgia, 2013, pp. 44–52.
- [17] S. Harispe, S. Ranwez, S. Janaqi, J. Montmain, The semantic measures library: assessing semantic similarity from knowledge representation analysis, in: E. Métais, M. Roche, M. Teisseire (Eds.), *Natural Language Processing and Information Systems*, Proc. of the 19th Intl. Conf. on Applications of Natural Language to Information Systems, LNCS, vol. 8455, Springer, Montpellier, France, 2014, pp. 254–257.
- [18] S. Harispe, S. Ranwez, S. Janaqi, J. Montmain, *Semantic similarity from natural language and ontology analysis*, Synthesis Lectures on Human Language Technologies, vol. 8, Morgan & Claypool Publishing, 2015.
- [19] S. Harispe, D. Sánchez, S. Ranwez, S. Janaqi, J. Montmain, A framework for unifying ontology-based semantic similarity measures: a study in the biomedical domain, *J. Biomed. Inform.* 48 (2014) 38–53.
- [20] F. Hill, R. Reichart, A. Korhonen, SimLex-999: Evaluating Semantic Models with (Genuine) Similarity Estimation, 2014. arXiv:1408.3456.
- [21] G. Hirst, D. St-Onge, Lexical chains as representations of context for the detection and correction of malapropisms, in: C. Fellbaum (Ed.), *WordNet: An Electronic Lexical Database*, Massachusetts Institute of Technology, 1998, pp. 305–332.
- [22] S.-L. Hsieh, W.-Y. Chang, C.-H. Chen, Y.-C. Weng, Semantic similarity measures in the biomedical domain by leveraging a web search engine, *IEEE J. Biomed. Health Inform.* 17 (4) (2013) 853–861.
- [23] J.J. Jiang, D.W. Conrath, Semantic similarity based on corpus statistics and lexical taxonomy, in: *Proceedings of the International Conference Research on Computational Linguistics (ROCLING X)*, 1997, pp. 19–33.
- [24] J.J. Lastra-Díaz, *Intrinsic Semantic Spaces for the representation of documents and semantic annotated data*. Master's thesis, Universidad Nacional de Educación a Distancia (UNED), Department of Computer Languages and Systems, 29 September 2014, <<http://e-spacio.uned.es/fez/view/bibliuned:master-ETSInformatica-LSI-Jlastra>>.
- [25] J.J. Lastra-Díaz, A. García-Serrano, System and method for the indexing and retrieval of semantically annotated data using an ontology-based information retrieval model, United States Patent and Trademark Office (USPTO) application US14/576,679, 19 December 2014.
- [26] J.J. Lastra-Díaz, A. García-Serrano, A novel family of IC-based similarity measures with a detailed experimental survey on WordNet, *Eng. Appl. Artif. Intell.* 46 (A) (2015) 140–153.
- [27] C. Leacock, M. Chodorow, Combining local context and WordNet similarity for word sense identification, in: C. Fellbaum (Ed.), *WordNet: An Electronic Lexical Database*, Massachusetts Institute of Technology, 1998, pp. 265–283.
- [28] J.H. Lee, M.H. Kim, Y.J. Lee, Information retrieval based on conceptual distance in is-a hierarchies, *J. Document.* 49 (2) (1993) 188–207.
- [29] Y. Li, Z.A. Bandar, D. McLean, An approach for measuring semantic similarity between words using multiple information sources, *IEEE Trans. Knowl. Data Eng.* 15 (4) (2003) 871–882.
- [30] D. Lin, An information-theoretic definition of similarity, in: *Proc. of the 15th International Conference on Machine Learning (ICML)*, vol. 98, Madison, WI, 1998, pp. 296–304.
- [31] L. Meng, J. Gu, A new model for measuring word sense similarity in WordNet, in: *Proc. of the 4th Intl. Conf. on Advanced Communication and Networking*, ASTL, vol. 14, 2012, pp. 18–23.
- [32] L. Meng, J. Gu, Z. Zhou, A new model of information content based on concept's topology for measuring semantic similarity in WordNet, *Int. J. Grid Distrib. Comput.* 5 (3) (2012) 81–93.
- [33] L. Meng, R. Huang, J. Gu, Measuring semantic similarity of word pairs using path and information content, *Int. J. Future Gener. Commun. Netw.* 7 (3) (2014) 183–194.
- [34] R. Mihalcea, C. Corley, C. Strapparava, Corpus-based and knowledge-based measures of text semantic similarity, *Proc. of the AAAI Conference on Artificial Intelligence*, vol. 1, AAAI Press, 2006, pp. 775–780.
- [35] G.A. Miller, W.G. Charles, Contextual correlates of semantic similarity, *Language Cognit. Process.* 6 (1) (1991) 1–28.
- [36] C. Orum, C.A. Joslyn, Valuations and Metrics on Partially Ordered Sets, March 2009. arXiv:0903.2679.
- [37] S. Patwardhan, S. Banerjee, T. Pedersen, Using measures of semantic relatedness for word sense disambiguation, in: A. Gelbukh (Ed.), *Proc. of the 4th Intl. Conf. on Computational Linguistics and Intelligent Text Processing (CICLing 2003)*, LNCS, vol. 2588, Springer, Mexico D.F., 2003, pp. 241–257.
- [38] S. Patwardhan, T. Pedersen, Using WordNet-based context vectors to estimate the semantic relatedness of concepts, in: *Proceedings of the EACL 2006 Workshop Making Sense of Sense-Bringing Computational Linguistics and Psycholinguistics Together*, vol. 1501, Trento, Italy, 2006, pp. 1–8.
- [39] T. Pedersen, WordNet-InfoContent-3.0.tar dataset repository, <https://www.researchgate.net/publication/273885902_WordNet-InfoContent-3.0.tar>.
- [40] T. Pedersen, Information content measures of semantic similarity perform better without sense-tagged text, in: *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT '10)*, Association for Computational Linguistics, Stroudsburg, PA, USA, 2010, pp. 329–332.
- [41] T. Pedersen, Measuring the Similarity and Relatedness of Concepts: a MICAI 2013 Tutorial, Tutorial presentation within the 12th Mexican Intl. Conf. on Artificial Intelligence, 25 November 2013, doi:<http://dx.doi.org/10.13140/RG.2.1.3025.6164>.
- [42] C. Pesquita, D. Faria, A.O. Falcao, P. Lord, F.M. Couto, Semantic similarity in biomedical ontologies, *PLoS Comput. Biol.* 5 (7) (2009) e1000443.
- [43] G. Pirró, A semantic similarity metric combining features and intrinsic information content, *Data Knowl. Eng.* 68 (11) (2009) 1289–1308.
- [44] G. Pirró, J. Euzenat, A feature and information theoretic framework for semantic similarity and relatedness, in: P.F. Patel-Schneider, Y. Pan, P. Hitzler, P. Mika, L. Zhang, J.Z. Pan, I. Horrocks, B. Glimm (Eds.), *Proc. of the 9th Intl. Semantic Web Conference, ISWC 2010*, LNCS, vol. 6496, Springer, Shanghai, China, 2010, pp. 615–630.

- [45] G. Pirró, N. Seco, Design, implementation and evaluation of a new semantic similarity metric combining features and intrinsic information content, in: R. Meersman, Z. Tari (Eds.), *On the Move to Meaningful Internet Systems: OTM 2008*, LNCS, vol. 5332, Springer, 2008, pp. 1271–1288.
- [46] R. Rada, H. Mili, E. Bicknell, M. Blettner, Development and application of a metric on semantic nets, *IEEE Trans. Syst. Man Cybernet.* 19 (1) (1989) 17–30.
- [47] P. Resnik, Using information content to evaluate semantic similarity in a taxonomy, in: *Proc. of the International Joint Conferences on Artificial Intelligence (IJCAI 1995)*, vol. 1, Montreal, Canada, 20 August 1995, pp. 448–453.
- [48] P. Resnik, Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language, *J. Artif. Intell. Res.* 11 (1999) 95–130.
- [49] H. Rubenstein, J.B. Goodenough, Contextual correlates of synonymy, *Commun. ACM* 8 (10) (1965) 627–633.
- [50] D. Sánchez, M. Batet, Semantic similarity estimation in the biomedical domain: an ontology-based information-theoretic perspective, *J. Biomed. Inform.* 44 (5) (2011) 749–759.
- [51] D. Sánchez, M. Batet, A new model to compute the information content of concepts from taxonomic knowledge, *Int. J. Semantic Web Inform. Syst.* 8 (2) (2012).
- [52] D. Sánchez, M. Batet, D. Isern, Ontology-based information content computation, *Knowl.-Based Syst.* 24 (2) (2011) 297–303.
- [53] D. Sánchez, M. Batet, D. Isern, A. Valls, Ontology-based semantic similarity: a new feature-based approach, *Expert Syst. Appl.* 39 (2012) 7718–7728.
- [54] K. Saruladha, G. Aghila, S. Raj, A survey of semantic similarity methods for ontology based information retrieval, in: *Proc. of the Second International Conference on Machine Learning and Computing (ICMLC 2010)*, IEEE, 2010, pp. 297–301.
- [55] A. Sebti, A.A. Barfroush, A new word sense similarity measure in WordNet, in: *Proc. of the Intl. Multiconference on Computer Science and Information Technology, IMCSIT 2008*, IEEE, 2008, pp. 369–373.
- [56] N. Seco, T. Veale, J. Hayes, An intrinsic information content metric for semantic similarity in WordNet, in: R. López de Mántaras, L. Saitta (Eds.), *Proceedings of the 16th European Conference on Artificial Intelligence (ECAI)*, vol. 16, IOS Press, Valencia, Spain, 2004, pp. 1089–1094.
- [57] A. Solé-Ribalta, D. Sánchez, M. Batet, F. Serratos, Towards the estimation of feature-based semantic similarity using multiple ontologies, *Knowl.-Based Syst.* 55 (2014) 101–113.
- [58] A. Tversky, Features of similarity, *Psychol. Rev.* 84 (4) (1977) 327–352.
- [59] J. Wang, H. Liu, H. Wang, A mapping-based tree similarity algorithm and its application to ontology alignment, *Knowl.-Based Syst.* 56 (2014) 97–107.
- [60] Z. Wu, M. Palmer, Verbs semantics and lexical selection, in: *Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics (ACL '94)*, Association for Computational Linguistics, Stroudsburg, PA, USA, 1994, pp. 133–138.
- [61] W. Yan, H. Liu, C. Zanni-Merk, D. Cavallucci, IngeniousTRIZ: an automatic ontology-based system for solving inventive problems, *Knowl.-Based Syst.* 75 (2015) 52–65.
- [62] Q. Yuan, Z. Yu, K. Wang, A new model of information content for measuring the semantic similarity between concepts, in: *Proc. of the Intl. Conference on Cloud Computing and Big Data (CloudCom-Asia 2013)*, IEEE Computer Society, 2013, pp. 141–146.
- [63] Z. Zhou, Y. Wang, J. Gu, A new model of information content for semantic similarity in WordNet, *Proc. of the Second Intl. Conference on Future Generation Communication and Networking Symposia (FGCN'S'08)*, vol. 3, IEEE, 2008, pp. 85–89.
- [64] Z. Zhou, Y. Wang, J. Gu, New model of semantic similarity measuring in WordNet, *Proc. of the 3rd Intl. Conference on Intelligent System and Knowledge Engineering (ISKE 2008)*, vol. 1, IEEE, 2008, pp. 256–261.
- [65] N. Zong, S. Nam, J.-H. Eom, J. Ahn, H. Joe, H.-G. Kim, Aligning ontologies with subsumption and equivalence relations in linked data, *Knowl.-Based Syst.* 76 (2015) 30–41.

Proofs of the propositions in the article:
“A new family of information content models with an
experimental survey on WordNet”

Juan J. Lastra-Díaz ¹ Ana García-Serrano ²

*NLP & IR Research Group
E.T.S.I. Informática - UNED
Universidad Nacional de Educación a Distancia
C/Juan del Rosal 16, 28040 Madrid*

Abstract

This brief notes provide the proofs of the two propositions supporting the axiomatic approach followed into the definition of the well-founded IC models described in the accompanying paper Lastra-Díaz and García-Serrano (2015). In order to make easier the reading of these notes, we have reproduced herein the whole section 4.1 titled “Preliminary concepts and notation”, including the proof of both propositions.

Key words: Semantic similarity, Intrinsic Information Content model, Ontology-based semantic similarity measures and distances, IC-based measures

1. Preliminary concepts and notation

For the sake of clarity, we use the lowercase letter p to denote a concept-valued probability function in a set of concepts C . On the other hand, the uppercase P is reserved to denote a probability measure, which is a set-valued function in the power set of the sample space. Finally, the conditional probability functions between concepts are denoted in lowercase by $p(c_i|c_j)$.

All the IC models proposed herein share the same computational structure, defined by the following three steps: (1) estimation of the edge-based conditional probabilities $p(c_i|c_j)$, (2) recovery of the concept-valued probability density function $p(c_i)$, and (3) computation of the node-based IC values using the standard definition $IC(c_i) = -\log_2(p(c_i))$. The only difference between the IC models is the method used to estimate the conditional probabilities. We call the new IC models *well-founded* because they are designed from first principles in order to satisfy the structural relationships of a discrete probability space and an Information Content model defined on this space.

In Jiang and Conrath (1997), the authors prove that their semantic distance $d_{J\&C}(c_1, c_2)$ is equivalent to the length of the shortest path between concepts c_1 and c_2 over a weighted graph derived from the taxonomy, and the edge weights

¹jlastra@invi.uned.es (corresponding author)

²agarcia@lsi.uned.es

are defined by (1). Despite the authors claiming that their distance is a metric on any type of taxonomy, nowadays, in Orum and Joslyn (2009) the authors prove that it is only true for the tree-like taxonomies, not for general taxonomies with multiple inheritance.

Every taxonomy $\mathcal{C} = (C, \leq_C, \Gamma)$ induces a graph $G = (E, V)$ in the usual manner, where every concept is a vertex of the graph, it means $V = C$, and there is an edge between each concept c_i and its direct parents, also called the lowest ancestors of c_i and denoted as $LA(c_i)$. The IC-based weighting function (1) allows us to introduce a shift of paradigm for the definition of the IC models, we move from a node-based IC computation model to an edge-based model.

$$\begin{aligned} w & : E \rightarrow \mathbb{R} \\ w(e_{ij}) & = -\log_2(p(c_i|c_j)) = IC(c_i) - IC(c_j) \\ E & = \{(c_i, c_j) \subset C \times C \mid c_j \in LA(c_i)\} \end{aligned} \tag{1}$$

Formally, a probability space is a triplet (Ω, \mathcal{F}, P) , where Ω is a non-empty set, called the space of outcomes or samples, \mathcal{F} is a σ -algebra that defines the collection of all possible events, where every event is defined as a subset of Ω , and finally, $P : \mathcal{F} \rightarrow \mathbb{R}$ is a probability measure. The formal definitions of the probability measures and probability spaces can be consulted in (Ash and Doléans-Dade, 2000, §1.2).

Definition 1 (Probability measure). *Given any non-empty set Ω and a collection \mathcal{F} of subsets on Ω , such that \mathcal{F} is a σ -algebra, then a set-valued function $P : \mathcal{F} \rightarrow \mathbb{R}$ is a probability measure if it satisfies the following axioms:*

1. $0 \leq P(A) \leq 1, \forall A \in \mathcal{F}$.
2. $P(\Omega) = 1, P(\emptyset) = 0$.
3. If $A = \{A_1, A_2, \dots, A_n\}$ is a family of disjoint subsets of \mathcal{F} , such that $\forall A_i, A_j \in A \implies A_i \cap A_j = \emptyset$, then:

$$P\left(\bigcup_{i \in I} A_i\right) = \sum_{i \in I} P(A_i)$$

If the space Ω is a countable set, and \mathcal{F} is the power set of the sample space Ω , denoted as 2^Ω , the triplet (Ω, \mathcal{F}, P) is called a discrete probability space (Ash and Doléans-Dade, 2000, §4.2). In our case, the space of samples Ω is discrete and is defined by the root concept Γ , such that $\Omega := \Gamma$, and the set \mathcal{F} is only the power set of the root concept Γ . Here, we are defining the root concept Γ as the universal set of the taxonomy, which follows that $\Gamma := C$. We note that in the last statement we are abusing the notation, because Γ is used to denote the root element of C and the sample space Ω at the same time.

We recall that the power set 2^Ω of any set Ω is a complete lattice when the inclusion relation \subseteq between subsets in Ω is used as an order relation. This fact is closely related to the relationship between the Jiang-Conrath distance and some types of metrics on lattices, such as we note in Lastra-Díaz (2014), and is detailed in a work on the metric properties of the Jiang-Conrath distance Orum and Joslyn (2009).

Given a taxonomy $\mathcal{C} = (C, \leq_C, \Gamma)$, where $L_C = \{c_k \in C \mid \nexists c_i \neq c_k, c_i \leq_C c_k\}$ is the set of leaves of the taxonomy, we can define a discrete probability space

(Γ, \mathcal{F}, P) on \mathcal{C} in the canonical manner as follows: (1) we define the root concept Γ as the universal sample space, (2) we define the set of leaf concepts $L_C \subset \mathcal{C}$ as the partition of disjoint sets of the sample space, such that $\Gamma := L_C$ by definition, (3) we define \mathcal{F} as the power set on \mathcal{C} , such that $c_i \subseteq c_j \iff c_i \leq_C c_j$, and finally, (4) we define a set-valued function $P : \mathcal{F} \rightarrow [0, 1]$ using a normalized leaf-valued function $p(c_k)$.

The triplet (Γ, \mathcal{F}, P) , as defined above, is a well-founded discrete probability space, a fact that we formalize in proposition 3, whose formal proof is omitted through lack of space. In order to get a well-founded probability space on any taxonomy, and derive the new family of intrinsic IC models from it. Below we provide a method to define any well-founded IC model based solely on the estimation of the conditional probabilities, which constitutes the core idea of this work.

Definition 2 (well-founded IC model). *Given a taxonomy of concepts $\mathcal{C} = (C, \leq_C, \Gamma)$, and an IC model defined by the function $IC : C \rightarrow \mathbb{R}^+ \cup \{0\}$, we call it a well-founded IC model if it can be written as $IC(c) = -\log_2(p(c))$ where $p(c)$ is a concept-valued function defined by (3), and the functions $p(c_i|c_j)$ are the conditional probabilities between any child concept c_i and its parent concepts c_j , which satisfy the edge-based property in (2).*

- (1) *Edge-based axiom.* The sum of conditional probabilities $p(c_i|c_j)$ of the children nodes c_i on any parent c_j node must be equal to 1, as defined in equation (2), where $LA(c_i)$ denotes the set of lowest ancestors (direct parents) of any concept c_i .

$$\sum_{\forall c_i|c_j \in LA(c_i)} p(c_i|c_j) = 1 \quad (2)$$

- (2) *Node-based axiom.* The probability $p(c_i)$ for each node c_i must be equal to the integration of the probabilities throughout the graph, starting from the root node, as defined in equation (3).

$$p : C \rightarrow [0, 1] \subset \mathbb{R}$$

$$p(c_i) = \begin{cases} 1 & , c_i = \Gamma \\ \sum_{\forall c_j \in LA(c_i)} p(c_j) p(c_i|c_j) & , c_i \neq \Gamma \end{cases} \quad (3)$$

- (3) *Leaf-based axiom.* The probabilities of the leaf concepts sum 1.

$$\sum_{c_k \in L_C} p(c_k) = 1 \quad (4)$$

The axioms (1) and (2) above allow us to define a family of well-founded intrinsic IC models based on the estimation of the conditional probabilities $p(c_i|c_j)$ for each edge of the taxonomy, such as is shown in table ???. In proposition 3, we show that given a taxonomy (C, \leq_C, Γ) , the leaf-based axiom (3) is a sufficient condition to get a well-founded probability space. In addition, we show in proposition 4 that axioms (1) and (2) of a well-founded IC model are

sufficient conditions to build a leaf-valued function $p : L_C \subset C \rightarrow [0, 1]$ that satisfies the IC model axiom (3). Thus, this last proposition proves that any well-founded IC model induces a well-founded probability space on any base taxonomy, and the whole system is supported by the structures derived from the conditional probabilities. We omit all the proofs herein by lack of space.

Proposition 3. *Be a taxonomy $\mathcal{C} = (C, \leq_C, \Gamma)$ defined by a partially ordered set (C, \leq_C) with a distinguished supreme element Γ , called the root, and L_C the set of leaves in C . If a set-valued positive function P is defined from the leaf-valued function p as follows:*

$$(1) \quad \begin{aligned} P : 2^\Gamma &\rightarrow [0, 1] \\ P(A) &= \sum_{c_k \in L_C \cap A} p(c_k) \end{aligned}$$

$$(2) \quad \begin{aligned} p : L_C \subset C &\rightarrow [0, 1] \\ \sum_{c_k \in L_C} p(c_k) &= 1 \end{aligned}$$

then the following facts are satisfied: (1) P is a probability measure, and (2) the triplet $(\Gamma, 2^\Gamma, P)$ is a probability space.

Proof.

If P is a probability measure then the triplet $(\Gamma, 2^\Gamma, P)$ is a probability space, because any power set on a discrete set is a σ -algebra. Thus we only need to prove that P is a probability measure.

Axiom 1. Because $\forall c_i \in C, p(c_i|c_j) \geq 0 \rightarrow p(c_i) \geq 0$, what follows that $\forall A \in 2^\Gamma, P(A) \geq 0$.

Axiom 2. By hypothesis, $P(\Gamma) = \sum_{c_k \in L_C \cap \Gamma} p(c_k) = \sum_{c_k \in L_C} p(c_k) = 1$
and $P(\emptyset) = \sum_{c_k \in L_C \cap \emptyset} p(c_k) = 0$.

Axiom 3. Now, we prove the countable additivity property.

(1) Given any taxonomy (C, \leq_C, Γ) , we can define a hierarchy of sets using the order relation \leq_C , such that $c_i \leq_C c_k \implies c_i \subseteq c_k$, Γ being the universal set for the inclusion relation \subseteq , such that $\forall c_i \in C \rightarrow c_i \subseteq \Gamma$.

(2) From (1) it follows that the leaves set $L_C = \{c_k \in C \mid \nexists c_i \neq c_k, c_i \leq_C c_k\}$ is a collection of pairwise disjoint sets.

(3) For L_C be a family of disjoint sets is not enough to be partition of Γ . Now, we must prove that L_C covers Γ . By definition $\forall c_k \in L_C$ is satisfied that $c_k \leq_C \Gamma$, so it follows that $L_C \subseteq \Gamma$. By other hand, we observe that for all the subsets $A \subseteq \Gamma$ there is always an element $c_k \in L_C$, such that $c_k \in A$, thus $\forall A \in 2^\Gamma \rightarrow A \cap L_C \neq \emptyset$, so it follows that $\Gamma \subseteq L_C$. Therefore, we prove that $\Gamma = L_C$, which proves that L_C is a partition of Γ .

(4) Because L_C is a partition of Γ , every set $A \in 2^\Gamma$ can be written as a finite union of elements in L_C , such that $A = \{c_k \in L_C\}_{k \in I}$.

(5) Be $c_1, c_2 \in C$ two disjoint sets, such that $c_1 \cap c_2 = \emptyset$. Then we have

$$P(c_1 \cup c_2) = \sum_{c_k \in L_C \cap (c_1 \cup c_2)} p(c_k) \quad (5)$$

(6) From $c_1 \cap c_2 = \emptyset$ so it follows that $L_C \cap (c_1 \cup c_2) = (L_C \cap c_1) \cup (L_C \cap c_2)$, equation (5) above takes the form below.

$$P(c_1 \cup c_2) = \sum_{c_k \in (L_C \cap c_1) \cup (L_C \cap c_2)} p(c_k) \quad (6)$$

$$P(c_1 \cup c_2) = \sum_{c_k \in (L_C \cap c_1)} p(c_k) + \sum_{c_k \in (L_C \cap c_2)} p(c_k) \quad (7)$$

$$P(c_1 \cup c_2) = P(c_1) + P(c_2) \quad (8)$$

(7) Finally, by hypothesis $c_1, c_2 \in C$ are any arbitrary pair of disjoint sets, thus, given an arbitrary family $A = \{A_1, A_2, \dots, A_n\}$ of pairwise disjoint subsets of Γ , using equation (8) above, we get the result below which proves the axiom 3 of a probability measure.

$$P\left(\bigcup_{i \in I} A_i\right) = \sum_{i \in I} P(A_i)$$

■

Proposition 4. *Be a taxonomy $\mathcal{C} = (C, \leq_C, \Gamma)$ and L_C the set of leaves in C . Given a concept-valued function p defined by*

$$p : C \rightarrow [0, 1]$$

$$p(c_i) = \begin{cases} 1 & , \text{ if } c_i = \Gamma \\ \sum_{\forall c_j \in LA_C(c_i)} p(c_i|c_j) p(c_j) & , \text{ otherwise} \end{cases}$$

then $P(L_C) = 1$, as given below:

$$P(L_C) = \sum_{c_k \in L_C} p(c_k) = 1$$

Proof. Our proof strategy is based on the definition of a bottom-up induction process, such that in each induction step all the leaf nodes are removed.

First, we will define two paired sequences of objects: (1) a sequence of ordered sets: $(C^0, \leq_{C^0}), (C^1, \leq_{C^1}), \dots, (C^n, \leq_{C^n})$, and (2) a sequence of probability measures: P^0, P^1, \dots, P^n .

The t-esim element of this sequence is grouped into a complex object called kernel and denoted by $K^t = ((C^t, \leq_{C^t}), P^t)$.

Next, we shall define an induction process such that the value of $P(L_C)$ is preserved for each step, it means that $P(L_C) = P^0(L_{C^0}) = P^1(L_{C^1}) = \dots = P^t(L_{C^t}) = \dots = P^n(L_{C^n}) = P(\Gamma) = 1$.

For each iteration, the kernel K^t induces a probability space on the reduced taxonomy, according to the construction defined in the proposition 3.

Induction step 0: The initial kernel K^0 is defined by the two items below:

$$\begin{cases} (C^0, \leq_{C^0}) = (C, \leq_C) \\ P^0(A) : 2^{C^0} \rightarrow [0, 1] \\ P^0(A) = \sum_{c_k \in L_{C^0} \cap A} p(c_k) \end{cases}$$

In step 0, we have that $L_C = L_{C^0}$, thus, $P(L_C) = P^0(L_{C^0})$, such as we want.

Induction step t-esim: for each iteration step, we remove the leaf nodes from K^t , resulting in a new kernel K^{t+1} defined below. Note that we are defining the order relation set $\leq_{C^{t+1}}$ as a correspondence over the Cartesian product $(C^t - L_{C^t})^2$.

$$\left\{ \begin{array}{l} (C^{t+1}, \leq_{C^{t+1}}) = (C^t - L_{C^t}, \leq_{C^{t+1}} = \{(c_i, c_j) \in (C^t - L_{C^t})^2 \mid c_i \leq_{C^t} c_j\}) \\ P^{t+1}(A) : 2^{C^{t+1}} \rightarrow [0, 1] \\ P^{t+1}(A) = \sum_{c_k \in L_{C^{t+1}} \cap A} p(c_k) \end{array} \right.$$

Next, we will prove that $P^0(L_{C^0}) = P^1(L_{C^1})$ and by induction, the iteration step above can be repeated n times until the resulting concept set be the root, it means $C^n = \Gamma$, thus, proving in this way that $P^n(L_{C^n}) = P(\Gamma) = 1$, and thus $P^0(L_{C^0}) = 1$, such as we wanted to prove.

Using the premise for the function $p(c_i)$, we can express $P^0(L_{C^0})$ as defined in equation (10). The term $LA_{C^t}(c_k)$ denotes the lowest ancestor set (parents) in t -*esim* iteration for any concept $c_k \in C^t$ within the ordered set (C^t, \leq_{C^t}) .

$$P^0(L_{C^0}) = \sum_{c_k \in L_{C^0}} p(c_k) \quad (9)$$

$$= \sum_{c_k \in L_{C^0}} \left(\sum_{\forall c_j \in LA_{C^0}(c_k)} p(c_k|c_j) p(c_j) \right) \quad (10)$$

Now, we can reverse the summation order in equation (10) to obtain the equation (11). Then, the sum runs over the union of all the $LA_{C^0}(c_k)$ sets, but $\bigcup_{c_k \in L_{C^0}} LA_{C^0}(c_k) = L_{C^1}$, therefore we obtain the equation (12).

$$P^0(L_{C^0}) = \sum_{\forall c_j \in \bigcup_{c_k \in L_{C^0}} LA_{C^0}(c_k)} \left(\sum_{c_k \in L_{C^0}} p(c_k|c_j) p(c_j) \right) \quad (11)$$

$$P^0(L_{C^0}) = \sum_{c_j \in L_{C^1}} \left(\sum_{c_k \in L_{C^0}} p(c_k|c_j) p(c_j) \right) \quad (12)$$

In equation (12), the inner sum runs over the leaf nodes L_{C^0} for a fixed parent node $c_j \in L_{C^1}$. Note that $\forall c_j \notin LA_{C^0}(c_k) \Rightarrow p(c_k|c_j) = 0$, what follows that the right inner sum over $c_k \in L_{C^0}$ is equal to $p(c_j)$, because $p(c_j)$ can be factorized from the inner product in equation (12), as shown in equation (13) below.

$$P^0(L_{C^0}) = \sum_{c_j \in L_{C^1}} \left(p(c_j) \underbrace{\left(\sum_{c_k \in L_{C^0}} p(c_k|c_j) \right)}_1 \right) = \sum_{c_j \in L_{C^1}} p(c_j) = P^1(L_{C^1}) \quad (13)$$

Finally, the sum of the conditional probabilities $p(c_k|c_j)$ in equation (13) is equal to 1 by definition. Therefore, we prove that $P^0(L_{C^0}) = P^1(L_{C^1})$, what follows that for each induction step $P^0(L_{C^0})$ is preserved and by induction $P^0(L_{C^0}) = P(\Gamma) = 1$, as we wanted to prove. ■

References

- Ash, R. B., Doléans-Dade, C. A., 2000. Probability & Measure Theory, 2nd Edition. Academic Press.
- Jiang, J. J., Conrath, D. W., 1997. Semantic similarity based on corpus statistics and lexical taxonomy. In: Proceedings of International Conference Research on Computational Linguistics (ROCLING X). pp. 19–33.
- Lastra-Díaz, J. J., 29 Sep. 2014. Intrinsic Semantic Spaces for the representation of documents and semantic annotated data. Master's thesis, Universidad Nacional de Educación a Distancia (UNED). Department of Computer Languages and Systems. <http://e-spacio.uned.es/fez/view/bibliuned:master-ETSIIinformatica-LSI-Jlastra>.
- Lastra-Díaz, J. J., García-Serrano, A., 2015. A new family of Information Content models with an experimental survey on WordNet. Knowledge-Based Systems Journal, doi: <http://dx.doi.org/10.1016/j.knosys.2015.08.019>.
- Orum, C., Joslyn, C. A., Mar. 2009. Valuations and Metrics on Partially Ordered Sets. arXiv:0903.2679.

Chapter 9

UNED Technical Report

This page intentionally left blank.



NLP and IR Research Group

ETSI Informática

Universidad Nacional de Educación a Distancia (UNED)

C/Juan del Rosal 16, 28040 Madrid (Spain)

A refinement of the well-founded Information Content models with a very detailed experimental survey on WordNet

Technical Report TR-2016-01

Juan J. Lastra-Díaz¹ Ana García-Serrano²

July 6, 2016

Cite this work as:

Lastra-Díaz, J. J., and García-Serrano, A. (2016). A refinement of the well-founded Information Content models with a very detailed experimental survey on WordNet. Technical Report TR-2016-01. NLP and IR Research Group. ETSI Informática. Universidad Nacional de Educación a Distancia (UNED). <http://e-spacio.uned.es/fez/view/bibliuned:DptoLSI-ETSI-Informes-Jlastra-refinement>

© 2016 The authors

¹ jlastra@invi.uned.es (corresponding author)

² agarcia@lsi.uned.es

This page is intentionally left in blank

A refinement of the well-founded Information Content models with a very detailed experimental survey on WordNet

Juan J. Lastra-Díaz Ana García-Serrano
(jlastra@invi.uned.es, agarcia@lsi.uned.es)

NLP and IR Research Group
ETSI Informática
Universidad Nacional de Educación a Distancia (UNED)
C/Juan del Rosal 16, 28040 Madrid (Spain)

July 11, 2016

Abstract

In a recent paper, we introduce a new family of Information Content (IC) models based on the estimation of the conditional probability between child and parent concepts. This work is encouraged by the finding of two drawbacks in the computational method of our aforementioned family of IC models, as well as other two gaps in the literature. First gap is that two of our cognitive IC models do not satisfy the axiom that constrains the sum of probabilities on the leaf nodes to be 1, whilst some ontologies with multiple inheritance could prevent the IC model satisfying the *growing monotonicity axiom* in concepts with multiple parents. Second gap is the lack of a complete and updated experimental survey including a pairwise statistical significance analysis between most IC models and ontology-based similarity measures. Finally a third gap is the lack of replication and confirmation of previous methods and results in most works. The latest two gaps are especially significant in the current state of the problem, in which there is no convincing winner within the family of intrinsic IC-based similarity measures and the performance margin is very narrow. In order to bridge the aforementioned gaps, this paper introduces the following contributions: (1) a refinement of our recent family of well-founded Information Content (IC) models; (2) eight new intrinsic IC models and one new corpus-based IC model; and (3) a very detailed experimental survey of ontology-based similarity measures and Information Content (IC) models on WordNet, including the evaluation and statistical significance analysis on the five most significant datasets of most ontology-based similarity measures and all WordNet-based IC models reported in the literature, with the only exception of the IC models recently introduced by [Harispe et al. \(2015a\)](#) and [Ben Aouicha et al. \(2016b\)](#). The evaluation is entirely based on a Java software library called HESML which has been developed by the authors in order to replicate all methods evaluated herein. The new IC models obtain rivaling results as regard the state-of-the-art methods and improve our previous models, whilst the experimental survey allows a detailed and conclusive image of the state of the problem to be drawn by setting the new state of the art and quantifying the main achievements of the last three decades.

Keywords: Intrinsic Information Content models, ontology-based semantic similarity measures, IC-based similarity measures, word similarity benchmark, semantic similarity, concept similarity model, experimental survey.

1 Introduction

The human similarity judgments between concepts underlie most of cognitive capabilities, such as categorization, memory, decision-making, and reasoning, as well as the use and discovery of analogies among others. For this reason, this problem has a lot of applications in Artificial Intelligence (AI) and many other related fields. The main research problem studied herein is the proposal of new Information Content (IC) models for ontology-based semantic similarity measures with the aim of estimating the degree of similarity between words as perceived by a human being. However, because of that the common ap-

proach to compute word similarity measures is to select the highest pairwise similarity value between the concept sets evoked by each word, our main research problem is closely related to the proposal of concept similarity models, whose aim is to estimate the degree of similarity between concepts instead of words. A concept similarity model is a function $sim : C \times C \rightarrow \mathbb{R}$ defined on a set of concepts which estimates the degree of similarity between concepts as perceived by a human being. The research into concept similarity models, so called in a broad sense as the human similarity judgment problem in cognitive sciences, has given rise to different strategies to tackle the problem of which the ontology-based simi-

Cite this work as: Lastra-Díaz, J. J., and García-Serrano, I. A. (2016). A refinement of the well-founded Information Content models with a very detailed experimental survey on WordNet. Technical Report TR-2016-01. NLP and IR Research Group. ETSI Informática. Universidad Nacional de Educación a Distancia (UNED).

<http://e-spacio.uned.es/fez/view/bibliuned:DptoLSI-ETSI-Infornes-Jlastra-refinement>

larity measures have proven to be the most successful of them.

The research into ontology-based semantic similarity measures is an old problem in AI and other related fields, such as cognitive psychology [Tversky \(1977\)](#), Natural Language Processing (NLP) and Information Retrieval (IR), [Rada et al. \(1989\)](#). A plethora of ontology-based similarity measures have been proposed in the literature, giving rise to a large set of applications in the fields of NLP, IR, bioengineering and genomics. For instance, [Lastra-Díaz \(2014\)](#) introduces an ontology-based IR model disclosed by [Lastra Díaz and García Serrano \(2014\)](#) which is based on the weighted Jiang-Conrath (J&C) distance introduced and evaluated in [Lastra-Díaz and García-Serrano \(2015b\)](#). [Patwardhan et al. \(2003\)](#) introduce a Word Sense Disambiguation (WSD) method based on the distributional hypothesis and the use of ontology-based similarity measures in order to select the closest evocated concept between a disambiguated word and its neighboring words. [Mihalcea et al. \(2006\)](#) propose a text similarity measure based on the combination of an Inverse Document Frequency (IDF) weighting scheme with any ontology-based similarity measure, which is evaluated in a Paraphrase Detection (PD) task, whilst [Fernando and Stevenson \(2008\)](#) propose a paraphrase detection method based on a quadratic form between Boolean occurrence vectors whose matrix is defined by any ontology-based similarity measure between words. In document clustering, [Song et al. \(2009\)](#) propose a genetic algorithm for text clustering based on a [Li et al. \(2003\)](#) similarity measure, whilst [Dagher and Fung \(2013\)](#) introduce a document clustering method based on a VSM model and a WordNet-based term expansion based on the [Jiang and Conrath \(1997\)](#) distance. [Liu et al. \(2009\)](#) introduce a method for the discovery of relevant WDSL-specified web services based on a WDSL similarity metric defined by the dot product between the provider and query vectors, whose weights are derived from the [Li et al. \(2003\)](#) similarity measure. [Martínez et al. \(2010\)](#) introduce a document anonymization method based on ontology-based similarity measures. [Cross and Hu \(2011\)](#) introduce a semantic alignment quality measure for the Ontology Alignment (OA) problem which relies on the difference between the similarity measure between the concepts in the base ontology and their image in the target ontology; and [Pirró and Talia \(2010\)](#) introduce an ontology mapping method based on a reformulation of the Jiang and Conrath (J&C) distance and the [Seco et al. \(2004\)](#) IC model, whilst [Jeong et al. \(2008\)](#) propose a framework for XML-schema matching based on ontology-based similarity measures. In [Oliva et al. \(2011\)](#), [Lee \(2011\)](#) and [Hadj Taieb et al. \(2015\)](#), the authors introduce different methods for sentence similarity based on ontology-based similarity measures. Other works use similarity measures for the extraction of domain ontologies from the Internet like [Wang and Zhou \(2009\)](#), or from text corpora like [Meijer et al. \(2014\)](#). [Montani et al. \(2015\)](#) propose an ontology-based process similarity metric for process mining that relies on the [Wu and Palmer \(1994\)](#) similarity measure. In the field of bioengineering, [Couto](#)

[et al. \(2007\)](#) introduce a reformulation of three classic IC-based similarity measures with the aim of computing similarity measures based on the Gene Ontology (GO), whilst [Chaves-González and Martínez-Gil \(2013\)](#) introduce a similarity-based evolutionary method for synonym recognition in the biomedical domain. Other specific similarity measures have been studied for biomedical text mining, such as [Pedersen et al. \(2007\)](#) and [Sánchez and Batet \(2011\)](#), as well as other genomics applications, such as protein function prediction [Pesquita et al. \(2009\)](#), [Couto and Pinto \(2013\)](#) and pathway prediction [Chiang et al. \(2008\)](#).

1.1 The context of our research

An ontology-based semantic similarity measure is a binary concept-valued function $sim : C \times C \rightarrow \mathbb{R}$ defined on a single-root taxonomy of concepts (C, \leq_C) which returns the degree of similarity between concepts as perceived by a human being. Modern research into the problem starts with the pioneering works by [Tversky \(1977\)](#) and [Rada et al. \(1989\)](#) in the fields of cognitive psychology and IR respectively. [Tversky \(1977\)](#) introduce a feature-based similarity measure which requires a representation of the concepts as feature sets, whilst [Rada et al. \(1989\)](#) introduce a semantic distance defined as the length of the shortest path between concepts in a taxonomy. The main drawback of the [Rada et al. \(1989\)](#) measure, as well as other similarity measures which use the length of the shortest path between concepts, is that all the edges in the taxonomy contribute to the overall distance with the same weight, the so-called *uniform weighting* problem. In order to bridge this latter gap, [Resnik \(1995\)](#) introduces the first similarity measure based on an Information Content (IC) model derived from corpus statistics, as well as the first method to compute an IC model, such as those proposed herein.

Every IC-based similarity measure needs a complementary concept-valued function, called the Information Content (IC) model. Given a taxonomy of concepts defined by a triplet $\mathcal{C} = ((C, \leq_C), \Gamma)$ where $\Gamma \in C$ is the supreme element called the root, an Information Content model is a function $IC : C \rightarrow \mathbb{R}^+ \cup \{0\}$, which represents an estimation of the information content for every concept, defined by $IC(c_i) = -\log_2(p(c_i))$, $p(c_i)$ being the occurrence probability of each concept $c_i \in C$. Every IC model must satisfy two further properties: (1) nullity in the root, such that $IC(\Gamma) = 0$, and (2) growing monotonicity from the root to the leaf concepts, such that $\forall c_i \leq_C c_j \Rightarrow IC(c_i) \geq IC(c_j)$. Once the IC-based measure is chosen, the IC model is mainly responsible for the definition of the notion of similarity and distance between concepts. Other works, such as [Pirró and Euzenat \(2010\)](#), have also proposed intrinsic IC models for semantic relatedness measures which rely on the whole set of semantic relationships encoded into an ontology.

The first known IC model is based on corpus statistics, which was introduced by [Resnik \(1995\)](#) and detailed in [Resnik \(1999\)](#). The main drawback of the corpus-based IC models is the difficulty in getting a well-balanced and disambiguated corpus for the estimation of the concept probabilities. To bridge this gap, [Seco et al. \(2004\)](#) intro-

duced the first intrinsic IC model in the literature, whose core hypothesis is that the IC models can be directly computed from intrinsic taxonomical features. Therefore, the development of new intrinsic IC-based similarity measures is divided into two subproblems: (1) the proposal of new intrinsic IC models, as in our work, and (2) the proposal of new IC-based similarity measures. In another recent work [Lastra-Díaz and García-Serrano \(2015a\)](#), we introduce a new family of intrinsic and corpus-based IC models called *well-founded IC models*, which is based on the proposal of different methods for the estimation of the conditional probabilities between child and parent concepts within a taxonomy. The main idea behind the new family of *well-founded IC models* is that any IC model should satisfy a set of axioms that algebraically link the conditional probabilities, probability function and IC model in order to define a well-founded probability space.

1.2 Motivation and hypotheses

The first motivation is the finding of two drawbacks in the algorithm to compute the family of *well-founded IC models* introduced in [Lastra-Díaz and García-Serrano \(2015a\)](#). First, the two intrinsic and cognitive IC models called *CondProbLogistic* and *CondProbCosine* do not satisfy the axiom that constrains the sum of probabilities on the leaf nodes to be 1. It is a consequence of the non-linear transformations applied to the conditional probabilities of these two models, a fact that was already mentioned in our aforementioned work. Second, in some cases, the ontologies with multiple inheritance could prevent the IC model satisfying the *growing monotonicity axiom* in concepts with multiple parents. This latest fact means that for some concept pairs $c_i, c_j \in C$, the constraint $c_i \leq_C c_j \Rightarrow IC(c_i) \geq IC(c_j)$ could be violated. In appendix B of our aforementioned work, we prove that the recovery algorithm based on the recursive formula in equation (3) is a sufficient condition for the sum of probabilities over the leaf nodes to be 1, what follows the underlying probability space is well-defined. However, if the taxonomy exhibits multiple inheritance, the probabilities $p(c_i)$ derived from equation (3) could be higher than the probability of any direct parent in some nodes with multiple parents, thus, leading to a violation of the aforementioned growing monotonicity axiom. Our main hypothesis is that the solution to these two drawbacks could lead us to an improvement in the performance of the family of well-founded IC models, in addition to fixing an algebraic inconsistency that moves the family of *well-founded IC model* away from their original design principles.

Second motivation of this work is the lack of an updated and exhaustive evaluation of ontology-based similarity measures and IC models in WordNet, as well as the lack of an exhaustive pairwise statistical significance analysis between them. In the literature, we find some out-of-date similarity benchmarks such as that reported by [Budanitsky and Hirst \(2001\)](#) and [Budanitsky and Hirst \(2006\)](#), and others, more recent but not exhaustive, such as [Hadj Taieb et al. \(2014b\)](#). The largest and most recent word similarity benchmarks in WordNet are

introduced by [Lastra-Díaz and García-Serrano \(2015a\)](#) and [Lastra-Díaz and García-Serrano \(2015b\)](#). However, not all of the hybrid IC-based similarity measures evaluated in the latest work have been previously evaluated with many IC models considered herein and the datasets introduced by [Miller and Charles \(1991\)](#), [Agirre et al. \(2009\)](#) and [Hill et al. \(2015\)](#). In addition, most ontology-based similarity measures have never been compared through a statistical significance analysis. Therefore, in the light of the results reported by [Lastra-Díaz and García-Serrano \(2015a\)](#), and in order to provide a conclusive image of the current state of the problem, we introduce herein a new and larger evaluation of IC models and ontology-based similarity measures than those available in the literature. This new evaluation is based on the most recently available datasets and our own software implementation of all the IC models and similarity measures evaluated herein, covering most developments from the pioneering works of [Rada et al. \(1989\)](#) and [Seco et al. \(2004\)](#).

Finally, the last motivation is the replication of previous methods and experiments. Most works introducing similarity measures or IC models during the last decade have only implemented or evaluated classic IC-based similarity measures, such as the Resnik, Lin and Jiang-Conrath measures, avoiding the replication of IC models and similarity measures introduced by other researchers. Some works have not included all the details of their methods, or the experimental setup to obtain the published results, thus, preventing their reproducibility. Most works have copied results published by others. This latest fact has prevented the valuable confirmation of previous methods and results reported in the literature, which is an essential feature of science. [Pedersen \(2008a\)](#), and subsequently [Fokkens et al. \(2013\)](#), warn of the need to reproduce and validate previous methods and results reported in the literature, a suggestion that we subscribe to in our aforementioned works, where we also warn of finding some contradictory results. This replication problem is especially significant in the current state of the problem, in which there is no convincing winner within the family of intrinsic IC-based similarity measures and the performance margin is very narrow, as concluded in our aforementioned works. In addition, [Pedersen \(2008a\)](#) also warns of the need of releasing the software developed for the evaluation of new methods and experiments reported in the literature with the aim of allowing their reproducibility. Following the suggestions from Pedersen, we introduce our new software library of ontology-based semantic similarity measures and IC models together with a set of reproducible experiments in a forthcoming paper, [Lastra-Díaz and García-Serrano \(2016\)](#).

The proposed refinements close the algebraic and algorithmic definition of the family of *well-founded IC models*, giving rise to research into further IC models within this family.

For the experimental survey, our main hypotheses are as follows:

H1. A group of recent IC-based similarity measures outperform the path-based similarity measures, as well

as the classic IC-based measures, but there is no statistically significant difference between them.

- H2.** There is no statistically significant difference in performance between most intrinsic IC models and the best performing corpus-based IC model defined as baseline, which is derived from the “*ic-treebank-add1.dat*” file in the Pedersen (2008b) dataset.
- H3.** A small set of the best performing intrinsic IC models outperform the best performing corpus-based IC model defined as baseline.
- H4.** The classic IC-based similarity measures proposed by Resnik, Jiang and Conrath, and Lin have been definitively outperformed by a small set of state-of-the-art IC-based similarity measures.
- H5.** The practical use of the current hybrid IC-based similarity measures that are based on the length of the shortest path is prevented by their high computational cost in comparison with the other methods with a similar performance.
- H6.** Most IC-based similarity measures perform better with a specific IC model.
- H7.** The state-of-the-art IC-based similarity measures outperform the best corpus-based similarity measures in the SimLex665 dataset.
- H8.** The proposed refinement into the computation method of the well-founded IC models could lead us to an improvement in their performance.

1.3 Research problem and contributions

The main aims of this paper are as follows. First, the proposal of a refinement into the four-step algorithm used to compute the family of well-founded IC models with the aim of eliminating the aforementioned drawbacks of the computational method introduced in our previous work, Lastra-Díaz and García-Serrano (2015a). Second, the proposal of eight new intrinsic IC models and one new corpus-based IC model in the new framework of our family of well-founded IC models. And third, the introduction of a new and very detailed experimental survey of IC models and ontology-based similarity measures on WordNet with a complete detailed statistical significance analysis between IC models and similarity measures, including the evaluation of most ontology-based similarity measures since the work of Rada et al. (1989) and all WordNet-based IC models reported in the literature, with the only exception of the IC models recently introduced by Harispe et al. (2015a) and Ben Aouicha et al. (2016b).

The refinement of the *well-founded IC models* allows a new family of IC models to be derived from the previous models introduced by Lastra-Díaz and García-Serrano (2015a), as well as three new strategies to compute the conditional probabilities. The new intrinsic IC models are called *CondProbRefHyponyms*, *CondProbRefUniform*, *CondProbRefLeaves*, *CondProbRefLogistic*, *CondProbRefCosine*, *CondProbRefLogisticLeaves*,

CondProbRefCosineLeaves and *CondProbRefLeavesSubsumersRatio*, whilst the new corpus-based IC model is called *CondProbRefCorpus*. The *CondProbRefLeavesSubsumersRatio* IC model is a reformulation of the Sánchez et al. (2011) IC model in the framework defined by our family of IC models.

The new experimental survey includes most of the intrinsic and corpus-based IC models evaluated in Lastra-Díaz and García-Serrano (2015a), as well as the nine new IC models introduced herein, one of the unexplored intrinsic IC models introduced by Blanchard et al. (2008), and most ontology-based similarity measures since the work by Rada et al. (1989). The word similarity benchmarks introduced herein include the five most significant datasets on the problem, as well as a very detailed pairwise statistical significance analysis between the IC models and ontology-based similarity measures. The benchmarks reported herein are, to the best of our knowledge, the largest experimental survey on intrinsic IC models and ontology-based similarity measures on WordNet reported in the literature, which is based on a same code implementation. We exactly reproduce the same experiments from Lastra-Díaz and García-Serrano (2015a), but with a much larger set of IC models and ontology-based similarity measures. Our experiments include a set of the hybrid IC-based similarity measures based on the length of the shortest path between concepts which were evaluated in Lastra-Díaz and García-Serrano (2015b) and subsequently discarded because of their high computational cost. The experimental survey includes 22 ontology-based similarity measures, 22 intrinsic IC models, and 3 corpus-based IC models.

The rest of the paper is structured as follows. Section 2 reviews the literature on concept similarity models. Section 3 summarizes the factual state of the art of the problem, whilst section 3.1 reviews the literature on intrinsic IC models. Section 4 introduces the proposed refinement in the well-founded IC models, as well as the new IC models derived from it. Section 5 describes the evaluation methodology and the results obtained. Section 6 introduces an in-depth discussion of the results. Last section presents our conclusions and future work. Finally, appendix groups the summary data tables and all raw data tables resulting from the evaluation.

2 Concept similarity models

This section makes a comparison between the concept and word similarity models proposed in the literature which we categorize as ontology-based and corpus-based similarity measures, and the most recent concept similarity models proposed in cognitive psychology. First, we compare the main strategies adopted to tackle the problem, and finally, we review the literature on corpus-based and ontology-based similarity measures.

2.1 Comparison of strategies

In the fields of NLP and IR, we find two different types of similarity models to estimate the degree of similarity between words: (1) ontology-based similarity measures as

Reference	Definition of the non IC-based similarity measures
	$sim_{Rada}(c_1, c_2) = 1 - \frac{1}{2}d_{Rada}(c_1, c_2)$
Rada et al. (1989)	$d_{Rada}(c_1, c_2) = len(c_1, c_2) = \min_{\forall \alpha \in Paths(c_1, c_2)} \left\{ \sum_{e_{ij} \in \alpha} 1 \right\}$
Wu and Palmer (1994)	$sim_{W\&P}(c_1, c_2) = \frac{2 \times depth(LCA(c_1, c_2))}{len(c_1, LCA(c_1, c_2)) + len(c_2, LCA(c_1, c_2)) + 2 \times depth(LCA(c_1, c_2))}$
Leacock and Chodorow (1998)	$sim_{L\&C}(c_1, c_2) = -\log\left(\frac{1 + len(c_1, c_2)}{2 \times maxdepth}\right)$
Li et al. (2003)	$sim_{Li_s3}(c_1, c_2) = e^{-\alpha * len(c_1, c_2)}, \quad \alpha^* = 0.25$
Li et al. (2003)	$sim_{Li_s4}(c_1, c_2) = e^{-\alpha * len(c_1, c_2)} \times \frac{e^{\beta * d} - e^{-\beta * d}}{e^{\beta * d} + e^{-\beta * d}}, \quad \alpha^* = 0.2 \quad \beta^* = 0.6$ $d = depth(LCA(c_1, c_2))$
Al-Mubaid and Nguyen (2009)	$d_{Mubaid}(c_1, c_2) = \log(1 + len(c_1, c_2) * (depthmax - depth(LCS(c_1, c_2))))$
Pedersen et al. (2007)	$sim_{Path}(c_1, c_2) = \frac{1}{1 + len(c_1, c_2)}$
Sánchez et al. (2012)	$dis_{S\&B}(c_1, c_2) = \log_2\left(1 + \frac{ \phi(c_1) \setminus \phi(c_2) + \phi(c_2) \setminus \phi(c_1) }{ \phi(c_1) \setminus \phi(c_2) + \phi(c_2) \setminus \phi(c_1) + \phi(c_1) \cap \phi(c_2) }\right)$ $\phi(a) = \{c \in C \mid a \leq c\}$ $sim_{T\&H}(c_1, c_2) = TermDepth(c_1, c_2) \times TermHypo(c_1, c_2)$ $TermDepth(c_1, c_2) = \frac{2 \times depth(c_1, c_2)}{depth(c_1) + depth(c_2)}$ $TermHypo(c_1, c_2) = \frac{2 \times SpecHypo(c_1, c_2)}{SpecHypo(c_1, c_2) + SpecHypo(c_2, c_1)}$ $SpecHypo(c_1, c_2) = 1 - \frac{\log(HypoValue(c))}{\log(HypoValue(root))}$ $HypoValue(c) = \sum_{c' \in HypoInc(c)} P(depth(c'))$ $P(depth(c')) = \frac{ \{c' \in C \mid depth(c') = depth(c)\} }{ C }$ $depth(c) = \text{length of the longest ascending path } c \rightarrow root$ $HypoInc(c) = \{c' \in C \mid c' \leq c\}$
Hadj Taieb et al. (2014b)	

Table 1: State-of-the-art non IC-based similarity measures evaluated in our experiments.

in our work, and (2) corpus-based similarity and relatedness measures. The ontology-based similarity measures are based on the definition of binary concept-valued similarity functions on “is-a” taxonomies, which have proven in Lastra-Díaz and García-Serrano (2015a) to be the best approximation to similarity human judgments on the noun subset of the SimLex dataset Hill et al. (2015), as being efficient, robust and easy to implement. However, the main drawback of the ontology-based similarity measures is the limited coverage of the ontologies and the cost and difficulties of building them. Other drawback of the ontology-based methods is the requirement of a single taxonomy that includes all the words to be compared, although this problem has given rise to the proposal of methods for the estimation of semantic similarity measures combining multiple ontologies, such as the general-purpose method introduced by Al-Mubaid and Nguyen (2009), the method for feature-based measures proposed by Solé-Ribalta et al. (2014) and the method for IC-based similarity measures proposed by Batet et al. (2014). On the other hand, the corpus-based similarity and relatedness measures mainly rely on the distributional hypothesis, and they are commonly based on the statistical co-occurrence between word contexts in large corpora, as a means of estimating the degree of similarity between words. The corpus-based measures “can confuse similarity with relatedness” (Li et al., 2015, §1). In addition, “it is commonly considered that distributional measures can only be used to capture semantic relatedness” (Harispe et al., 2015b, §2.5.2), and “they have traditionally performed poorly when compared to WordNet-based measures” (Mohammad and Hirst, 2012, p.1). This latter fact is confirmed by the recent compar-

isons between ontology-based and corpus-based similarity measures reported by (Banjade et al., 2015, Table 1) and Le and Fokkens (2015), as well as our benchmarks in (Lastra-Díaz and García-Serrano, 2015a, §6.4). It is worth to note that the ontology-based similarity measures use an explicitly defined concept similarity model with the aim of estimating the degree of similarity between words whose specific meaning (evocated concept) is unknown, whilst the corpus-based measures use the occurrence of the words in a specific context, whose meaning (concept) is implicitly defined by the context.

Finally, the research into the similarity judgments problem in cognitive psychology derives from the pioneering work of Tversky (1977). The research into the field of IR has focused on the proposal of a plethora of symmetric and contextless similarity measures guided by experimental evaluation. On the contrary, the research into cognitive sciences has followed a parallel line more focused on the definition of theoretical models capable of explaining several non-metric phenomena in the human similarity judgments described by Tversky (1977) and Pothos et al. (2015), such as: (1) asymmetry or non-commutativity, (2) context dependency and (3) the conjunction fallacy. The most recent cognitive similarity model is introduced by Pothos et al. (2013) and Pothos and Trueblood (2015), being inspired by a quantum probability approach for cognition proposed by Busemeyer and Bruza (2012), whose non-commutative nature allows the representation of different non-metric phenomena. However, the quantum probability similarity model has not yet been experimentally evaluated.

2.2 Corpus-based measures

Many corpus-based similarity or relatedness measures are based on concept-based resources, such as Wikipedia. For instance, [Strube and Ponzetto \(2006\)](#) introduce WikiRelate, a method for computing the semantic relatedness between words based on a graph derived from Wikipedia. WikiRelate extracts the Wikipedia pages associated to each input word and builds a taxonomy of categories by merging the categories that the pages belong to. Finally, WikiRelate uses standard path-based and IC-based similarity measures on the recovered taxonomy in order to compute the relatedness measure between words. We can interpret WikiRelate as a two-stage method based on the combination of a taxonomy recovering method, such as the method recently proposed by [Ben Aouicha et al. \(2016a\)](#), with any standard ontology-based similarity measure. [Gabrilovich and Markovitch \(2007\)](#) introduce a semantic relatedness method for word and documents, called ESA, which represents the meaning of a word or text as a weighted vector of Wikipedia concepts (articles); whilst [Agirre et al. \(2009\)](#) introduce several distributional relatedness measures based on a vector space model trained on a large Web corpus, which favourably compare with a large set of ontology-based similarity measures on WordNet.

On the other hand, another very active line of research in corpus-based similarity measures is the proposal for hybrid concept-based distributional measures, which integrate knowledge bases (KBs) or explicit “is-a” semantic networks in order to overcome the lack of well-defined semantic knowledge. For instance, [Patwardhan and Pedersen \(2006\)](#) introduce a similarity and relatedness measure which relies on the gloss vector overlapping between the extended WordNet gloss vectors of two input concepts. [Mohammad and Hirst \(2006\)](#) introduce a hybrid distributional measure which relies on the cosine function and the concept-based conditional probabilities for the words derived from the Roget’s thesaurus. [Alvarez and Lim \(2007\)](#) propose a hybrid distributional similarity measure that relies on the product of two taxonomical WordNet-based functions with a gloss overlapping factor by using “is-a” and “part-of” relationships, whilst [Li et al. \(2015\)](#) introduce another hybrid distributional measure whose core idea is that the similarity computation relies on truly “is-a” relationships, which are derived from a very large web corpus by using an automatic method based on syntactic rules.

Other family of relatedness measures are based on random walks on weighted graphs derived from different knowledges sources, such as Wikipedia and WordNet. For instance, [Hughes and Ramage \(2007\)](#) propose a semantic relatedness measure between word pairs which is based on a random walk using Personalized PageRank on a weighted graph derived from WordNet and corpus statistics, whilst [Yeh et al. \(2009\)](#) extend their previous work on semantic relatedness measures based on random walks to Wikipedia, and [Ramage et al. \(2009\)](#) propose a corpus-based measure based on a random walk on WordNet with the aim of estimating the semantic similarity between text fragments. Finally, [Yazdani and Popescu-Belis \(2013\)](#) propose a method for estimating the se-

mantic relatedness between concepts based on a random walk approach on a Wikipedia concept network with two link types: the hypertext links between Wikipedia articles (concepts), and the lexical similarity between them defined by the cosine score between the vectors representing each article.

Another growing research trend on corpus-based semantic similarity and relatedness measures is the development of word embeddings, such as those proposed by [Mikolov et al. \(2013\)](#), [Pennington et al. \(2014\)](#) and [Suzuki and Nagata \(2015\)](#), whose core idea is the learning of a vector representation (embedding) for large vocabularies, such that the Euclidean distance between word vectors reflects their semantic similarity. Most word embeddings use a large corpora in their learning process, thus, they are a subfamily of the corpus-based methods. The word embedding methods commonly use complex machine learning algorithms, which are time-consuming and hard to reproduce. However, once the vector representations are computed, their evaluation mainly depends on the dimensionality of the vector space, thus, they can be very efficient for large vocabularies and low dimensionality.

2.3 Ontology-based similarity measures

In two recent works, [Lastra-Díaz and García-Serrano \(2015b\)](#) and [Lastra-Díaz and García-Serrano \(2015a\)](#), we provide a very detailed review of the current ontology-based semantic measures, thus, we only provide herein a categorization in order to introduce the similarity measures that will be evaluated in our experiments. For a more in-depth review of the topic, we refer the reader to our aforementioned works, especially the former, and the recent book by [Harispe et al. \(2015b\)](#).

We categorize the current ontology-based semantic measures into four subfamilies as follows: (1) edge-counting similarity measures, the so called path-based measures, whose core idea is the use of the length of the shortest path between concepts as an estimation of their degree of similarity, such as the pioneering work of [Rada et al. \(1989\)](#) and the subsequent works of [Wu and Palmer \(1994\)](#), [Leacock and Chodorow \(1998\)](#), [Hirst and St-Onge \(1998\)](#), [Pedersen et al. \(2007\)](#) and [Al-Mubaid and Nguyen \(2009\)](#); (2) IC-based similarity measures whose core idea is the use of an Information Content (IC) model, such as the pioneering work of [Resnik \(1995\)](#), and the measures proposed by [Jiang and Conrath \(1997\)](#) and [Lin \(1998\)](#); (3) feature-based measures, whose core idea is the use of set-theory operators between the feature sets of the concepts, such as the pioneering work of [Tversky \(1977\)](#), and more recently [Sánchez et al. \(2012\)](#), whose core idea is the use of the overlapping of ancestor sets as an estimation of the overlapping between the unknown feature sets of the concepts; and finally, (4) other similarity measures that cannot be directly categorized into any previous family, which are based on taxonomical features derived from set-theory operators [Batet et al. \(2011\)](#), or novel contributions of the hyponym set [Hadj Taieb et al. \(2014b\)](#). Out of our previous categorization, it was also worth mentioning some proposals of *aggregated similarity measures*, such as [Martinez-Gil \(2016\)](#), whose key

Classic IC-based similarity measures	
Resnik (1995)	$sim_{Resnik}(c_1, c_2) = IC(MICA(c_1, c_2))$
Jiang and Conrath (1997)	$d_{J\&C}(c_1, c_2) = IC(c_1) + IC(c_2) - 2IC(MICA(c_1, c_2))$ $sim_{J\&C}(c_1, c_2) = 1 - \frac{1}{2}d_{J\&C}(c_1, c_2)$
Lin (1998)	$sim_{Lin}(c_1, c_2) = \frac{2IC(MICA(c_1, c_2))}{IC(c_1) + IC(c_2)}$
IC-based reformulations of the Tversky similarity measure	
Pirró and Seco (2008)	$sim_{P\&S}(c_1, c_2) = \begin{cases} \frac{3IC(MICA(c_1, c_2))}{-IC(c_1) - IC(c_2)} & , \text{ if } c_1 \neq c_2 \\ 1 & , \text{ if } c_1 = c_2 \end{cases}$
Monotone transformations of classic IC-based similarity measures	
Pirró and Euzenat (2010)	$sim_{FaITH}(c_1, c_2) = \frac{IC(MICA(c_1, c_2))}{IC(c_1) + IC(c_2) - IC(MICA(c_1, c_2))}$
Meng and Gu (2012)	$sim_{Meng}(c_1, c_2) = e^{sim_{Lin}(c_1, c_2)} - 1 = e^{\frac{2IC(MICA(c_1, c_2))}{IC(c_1) + IC(c_2)}} - 1$
Garla and Brandt (2012)	$sim_{path_IC}(c_1, c_2) = \frac{1}{1 + d_{J\&C}(c_1, c_2)}$
Lastra-Díaz and García-Serrano (2015b)	$sim_{cosJ\&C}(c_1, c_2) = 1 - \cos\left(\frac{\pi}{2} \left(1 - \frac{d_{J\&C}(c_1, c_2)}{2 * max_{d_{J\&C}}}\right)\right)$ $max_{d_{J\&C}} = \max_{c \in Leaves(C)} \{IC(c)\}$
Hybrid IC-based similarity measures based on the shortest path length	
Li et al. (2003)	$sim_{Li_s9}(c_1, c_2) = sim_{Li_s4}(c_1, c_2) * \frac{e^{\lambda * IC} - e^{-\lambda * IC}}{e^{\lambda * IC} + e^{-\lambda * IC}}, \lambda^* = 0.4$ $IC = MICA(c_1, c_2)$
Zhou et al. (2008b)	$sim_{Zh}(c_1, c_2) = 1 - k \times \left(\frac{\log(len(c_1, c_2) + 1)}{\log(2 * max_{e \in T} \{depth(c)\} - 1)} \right)$ $-\frac{1}{2}(1 - k) \times d_{J\&C}(c_1, c_2) \quad k^* = \frac{1}{2} \text{ by default}$
Meng et al. (2014)	$sim_{Meng2014}(c_1, c_2) = sim_{Lin}(c_1, c_2) \left(\frac{1 - e^{-k * len(c_1, c_2)}}{e^{-k * len(c_1, c_2)}} \right), k^* = 0.08$ $sim_{Gao}(c_1, c_2) = e^{-\alpha L(c_1, c_2)}, \alpha^* = 0.15 \text{ and } \beta^* = 2.05$
Gao et al. (2015)	$L(c_1, c_2) = wt(c_1, c_2) * len(c_1, c_2)$ $wt = \begin{cases} \left(\frac{1 + IC(MICA(c_1, c_2))}{IC(MICA(c_1, c_2))} \right)^\beta & , IC(MICA(c_1, c_2)) \geq 1 \\ 2^\beta & , 1 > IC(MICA(c_1, c_2)) \geq 0 \end{cases}$
Lastra-Díaz and García-Serrano (2015b)	$sim_{coswJ\&C}(c_1, c_2) = 1 - \cos\left(\frac{\pi}{2} \left(1 - \frac{d_{wJ\&C}(c_1, c_2)}{2 * max_{d_{J\&C}}}\right)\right)$ $d_{wJ\&C}(c_1, c_2) = \min_{\forall \alpha \in Paths(c_1, c_2)} \left\{ \sum_{e_{ij} \in \alpha} w(e_{ij}) \right\}$ $w(e_{ij}) = \begin{cases} -\log_2(p(c_i c_j)) & , \text{ if } p(c_i c_j) \text{ are known} \\ IC(c_i) - IC(c_j) & , \text{ otherwise} \end{cases}$

Table 2: Definition of the state-of-the-art IC-based similarity measures evaluated in our experiments.

feature is the merging of multiple ontology-based similarity measures in order to produce a final similarity judgement.

In addition to the four subfamilies of ontology-based similarity measures aforementioned above, we categorize the family of IC-based similarity measures into the following four subgroups, as shown in table 2: (1) the first group of classic IC-based measures made up of the similarity measures introduced by Resnik (1995), Jiang and Conrath (1997) and Lin (1998); (2) a second group that we call hybrid or path-based IC-based similarity measures, which is defined by those measures that make up an IC model with any function based on the length of the shortest path between concepts, such as the pioneering work of Li et al. (2003), and other subsequent works such as Zhou et al. (2008a), Meng et al. (2014), Gao et al. (2015), and the two weighted IC-based similarity measures introduced by Lastra-Díaz and García-Serrano

(2015b); (3) a third group that is based on any reformulation strategy between different approaches, such as the IC-based reformulations of the Tversky measure in Pirró (2009) and Pirró and Euzenat (2010), as well as the IC-based reformulation of most edge-counting methods introduced by Sánchez and Batet (2011); and finally, (4) a fourth group that is based on a monotone transformation of any classic IC-based similarity measure, such as the exponential-like scaling of the Lin (1998) measure introduced by Meng and Gu (2012), the reciprocal of the J&C distance introduced by Garla and Brandt (2012), and another cosine-based normalization of the J&C distance introduced by Lastra-Díaz and García-Serrano (2015b). In addition, we show herein that the FaITH similarity measure introduced by Pirró and Euzenat (2010) is also a monotone transformation of the Lin (1998) similarity measure, despite its initial design being based on a reformulation of the Tversky (1977) measure. Table 3 shows

Rada et al. (1989) similarity measure and its monotone transformations	
	$sim_{Rada}(c_1, c_2) = 1 - \frac{1}{2}d_{Rada}(c_1, c_2)$
Rada et al. (1989)	$d_{Rada}(c_1, c_2) = len(c_1, c_2) = \min_{\forall \alpha \in Paths(c_1, c_2)} \left\{ \sum_{e_{ij} \in \alpha} 1 \right\}$
	$sim_{L\&C}(c_1, c_2) = -\log\left(\frac{1+len(c_1, c_2)}{2 \times maxdepth}\right)$
Leacock and Chodorow (1998)	Factorization: $sim_{L\&C}(c_1, c_2) = \varphi(x) \circ sim_{Rada}(c_1, c_2)$ $\varphi(x) = -\log\left(\frac{3-2x}{2 \times maxdepth}\right)$
	$sim_{Li_s3}(c_1, c_2) = e^{-\alpha^* len(c_1, c_2)}, \quad \alpha^* = 0.25$
Li et al. (2003)	Factorization: $sim_{Li_s3}(c_1, c_2) = \varphi(x) \circ sim_{Rada}(c_1, c_2)$ $\varphi(x) = e^{2\alpha^*(x-1)} \xrightarrow{\alpha^*=0.25} \varphi^*(x) = e^{\frac{(x-1)}{2}}$ $sim_{Path}(c_1, c_2) = \frac{1}{1+len(c_1, c_2)}$
Pedersen et al. (2007)	Factorization: $sim_{Path}(c_1, c_2) = \varphi(x) \circ sim_{Rada}(c_1, c_2)$ $\varphi(x) = \frac{1}{3-2x}$
Lin (1997) similarity measure and its monotone transformations	
Lin (1998)	$sim_{Lin}(c_1, c_2) = \frac{2IC(MICA(c_1, c_2))}{IC(c_1)+IC(c_2)}$
	$sim_{FaITH}(c_1, c_2) = \frac{IC(MICA(c_1, c_2))}{IC(c_1)+IC(c_2)-IC(MICA(c_1, c_2))}$
Pirr6 and Euzenat (2010)	Factorization: $sim_{FaITH}(c_1, c_2) = \varphi(x) \circ sim_{Lin}(c_1, c_2)$ $\varphi(x) = \frac{x}{2-x}$ $sim_{Meng}(c_1, c_2) = e^{sim_{Lin}(c_1, c_2)} - 1$
Meng and Gu (2012)	Factorization: $sim_{Meng}(c_1, c_2) = \varphi(x) \circ sim_{Lin}(c_1, c_2)$ $\varphi(x) = e^x - 1$
Jiang and Conrath (1997) similarity measure and its monotone transformations	
Jiang and Conrath (1997)	$d_{J\&C}(c_1, c_2) = IC(c_1) + IC(c_2) - 2IC(MICA(c_1, c_2))$
	$sim_{J\&C}(c_1, c_2) = 1 - \frac{1}{2}d_{J\&C}(c_1, c_2)$
	$sim_{path_IC}(c_1, c_2) = \frac{1}{1+d_{J\&C}(c_1, c_2)}$
Garla and Brandt (2012)	Factorization: $sim_{path_IC}(c_1, c_2) = \varphi(x) \circ sim_{J\&C}(c_1, c_2)$ $\varphi(x) = 3 - 2x$ $sim_{cosJ\&C}(c_1, c_2) = 1 - \cos\left(\frac{\pi}{2} \left(1 - \frac{d_{J\&C}(c_1, c_2)}{2 * max_{d_{J\&C}}}\right)\right)$ $max_{d_{J\&C}} = \max_{c \in Leaves(C)} \{IC(c)\}$
Lastra-Díaz and García-Serrano (2015b)	Factorization: $sim_{cosJ\&C}(c_1, c_2) = \varphi(x) \circ \phi(t) \circ sim_{J\&C}(c_1, c_2)$ $\varphi(x) = 1 - \cos\left(\frac{\pi}{2}x\right)$ $\phi(t) = 1 - \frac{1-t}{max_{d_{J\&C}}}$, normalization function

Table 3: Equivalence classes of similarity measures induced by any monotone transformation from any classic similarity measure.

the monotonicity relationships between most IC-based similarity measures which have been experimentally confirmed in our evaluation. For the sake of completeness of our experimental survey, we also evaluate herein all non IC-based similarity measures shown in table 1, despite the present work is focused on new IC models and their evaluation with the state-of-the-art IC-based similarity measures shown in table 2.

Finally, Stanchev (2014) introduces a similarity graph from WordNet with the aim of computing the similarity between words. In addition to the taxonomical structure from WordNet, the graph uses the definition and examples of use of the WordNet concepts as evidence on the relationships between concepts. The similarity graph is defined by a collection of oriented edges with asymmetric weights, in which the weights between parent and child concepts encode the probability that a user interested in the source node of an edge is also interested in the concept associated to the destination node. The similarity measure is defined as the product of the edge weights throughout the path between the word nodes. Despite some weights being defined in an arbitrary way, the method obtains outstanding results in the Miller and Charles (1991) dataset, and introduces for the first time an asymmetrical path-based method founded on probability theory. We note that the similarity measure introduced by Stanchev is closely related to our weighted J&C distance, denoted by $dwJ\&C$ in table 2, as our measure matches the logarithm of the product of conditional probabilities between the word nodes. However, the basic form of the $dwJ\&C$ distance does not integrate the word nodes into the WordNet taxonomy and the weights are symmetric, the edge weights being the logarithm of the conditional probabilities.

2.4 Summary and positioning

In summary, the ontology-based similarity measures are efficient, easy to implement and more accurate than the corpus-based methods, whilst the corpus-based measures offer a broader lexical coverage at the expense of a high complexity and computational cost, as well as the difficulties to obtain well-balanced learning corpus. However, the corpus-based relatedness measures based on word embeddings combine the broad coverage of the corpus-based methods with an efficient evaluation method in operation mode. On the other hand, unlike the theoretical models developed in cognitive psychology which have not yet evaluated, the ontology-based similarity measures have been successfully evaluated in many human similarity benchmarks, and they have contributed to the development of a large set of applications. For these reasons, we are focusing our research effort on the development of new IC models and ontology-based similarity measures.

3 State of the art

This section summarizes the current factual state of the art on ontology-based similarity measures and IC models and review the related work on IC models.

The state of the art in ontology-based similarity measures is defined by the family of intrinsic IC-based measures, which are defined by the combination of one specific IC-based similarity measure with any intrinsic IC model. More specifically, our cosine-normalized Jiang-Conrath ($cosJ\&C$) similarity measure is currently the best performing ontology-based similarity measure according to the evaluation on the five most significant datasets reported in (Lastra-Díaz and García-Serrano, 2015a, table 6). However, in this latest work we did not evaluate other hybrid IC-based measures that obtained state-of-the-art results in Lastra-Díaz and García-Serrano (2015b), such as our hybrid measure $coswJ\&C$ and the Zhou et al. (2008b) similarity measure. Likewise, the $cosJ\&C$ similarity measure is the only measure that obtains a statistically significant higher performance than the baseline, (Lastra-Díaz and García-Serrano, 2015a, fig.3). However, we also prove that there is no statistically significant difference between the $cosJ\&C$ similarity measure and those introduced by Meng and Gu (2012) and Pirró and Euzenat (2010).

The outperformance of the IC-based similarity measures is supported by several recent WordNet-based benchmarks, such as Lastra-Díaz and García-Serrano (2015a), Lastra-Díaz and García-Serrano (2015b) and Hadj Taieb et al. (2014b), as well as other older ones, such as Budanitsky and Hirst (2006), Pirró (2009) and Sánchez et al. (2011). Another benchmark in bioengineering introduced by Garla and Brandt (2012) also confirms the outperformance of an intrinsic IC-based similarity measure derived from the reciprocal of the J&C distance. Likewise, McInnes and Pedersen (2013) prove the outperformance of the classic IC-based similarity measures over the path-based measures and gloss-based relatedness measures in a WSD benchmark in bioengineering, but it is also proven that there is no a statistically significant difference between a corpus-based IC model and the intrinsic IC model introduced by Sánchez et al. (2011). This latest conclusion on the debate between intrinsic and corpus-based IC models is endorsed in a more conclusive manner by the recent benchmarks in our aforementioned works.

In our aforementioned works, we conclusively prove several significant facts on the state of the art of IC models as follows. First, contrary to what the research community thought, most corpus-based IC models derived from the unexplored “*.add1” set of WordNet-based frequency files in Pedersen (2008b) rival the state-of-the-art intrinsic IC models, (Lastra-Díaz and García-Serrano, 2015b, table 6). Second, the best performing IC model on average is the Seco et al. (2004) IC model, (Lastra-Díaz and García-Serrano, 2015a, table 5). Third, there is no a statistical significant difference between most state-of-the-art intrinsic IC models, as well as between most intrinsic IC models and the baseline IC model defined by a corpus-based IC model derived from the “ic-treebank-add1.dat” file in the aforementioned Pedersen dataset, (Lastra-Díaz and García-Serrano, 2015a, fig.2). And finally, the Sánchez and Batet (2012) IC model is the only one that obtains a statistically significant higher performance than the corpus-based IC model defined as

IC models	Definition
Resnik (1999)	$IC_{Resnik} = -\log_2(\hat{p}(c_i))$ $\hat{p}(c_i) = \frac{f(c_i)}{N} = \frac{f(c_i)}{f(\Gamma)}$ $f(c_i) = TF(c_i) + IF(c_i) = TF(c_i) + \sum_{\forall c_j c_i \in LA(c_j)} f(c_j)$
Seco et al. (2004)	$IC_{Seco}(c) = 1 - \frac{\log(Hypo(c) +1)}{\log(max_nodes)}$
Zhou et al. (2008a)	$IC_{Zhou}(c) = k \left(1 - \frac{\log(Hypo(c) +1)}{\log(max_nodes)}\right) + (1-k) \frac{\log(depth(c))}{\log(depth_{max})}$ $k^* = \frac{1}{2}$
Blanchard et al. (2008)	$IC_g(c_i) = -\log_2\left(\frac{ SubsumedLeaves(c_i) }{maxLeaves}\right)$ $SubsumedLeaves(c_i) = \{c_j \in C \mid c_j \leq_C c_i \wedge c_j \text{ is leaf}\}$
Sánchez et al. (2011)	$IC_{Sánchez2011}(c_i) = -\log_2\left(\frac{\frac{ Leaves(c_i) }{ subsumers(c_i) } + 1}{maxLeaves + 1}\right)$ $\begin{cases} Leaves(c_i) = \{c_j \in C \mid (c_j \leq_C c_i \wedge c_j \neq c_i) \wedge c_j \text{ is leaf}\} \\ subsumers(c_i) = \{c_j \in C \mid c_i \leq_C c_j\} \end{cases}$
Sánchez and Batet (2012)	$IC_{Sánchez2012}(c) = -\log_2\left(\frac{commonness(c)}{commonness(root)}\right)$ $\begin{cases} commonness(c) = \frac{1}{ Subsumers(c) }, c \text{ leaf} \\ commonness(c) = \sum_{\forall l \mid l \text{ is leaf and } l < c} commonness(l), c \text{ not leaf} \end{cases}$
Meng et al. (2012)	$IC_{Meng}(c) = \frac{\log(depth(c))}{\log(depth_{max})} \times \left(1 - \frac{\log\left(1 + \sum_{a \in Hypo(c)} \frac{1}{depth(a)}\right)}{\log(Node_{max})}\right)$
Yuan et al. (2013)	$IC_{Yuan}(c) = f_{depth}(c) (1 - f_{leaves}(c)) + f_{hyper}(c)$ $\begin{cases} f_{depth}(c) = \frac{\log(depth(c))}{\log(depth_{max})} \\ f_{leaves}(c) = \frac{\log(Leaves(c) +1)}{\log(Leaves_{max}+1)} \\ f_{hyper}(c) = \frac{\log(Hyper(c) +1)}{\log(Node_{max})} \end{cases}$
Hadj Taieb et al. (2014a)	$IC_{Taieb}(c) = \left(\sum_{a \in HyperInc(c)} Score(a)\right) \times AvgDepth(c)$ $AvgDepth(c) = \frac{1}{ HyperInc(c) } \times \sum_{c' \in HyperInc(c)} depth(c')$ $Score(c) = \left(\sum_{c' \in DirectHyper(c)} \frac{depth(c')}{ HypoInc(c') }\right) \times HypoInc(c) $ $HypoInc(c) = \{a \in C \mid a \leq c\} \quad HyperInc(c) = \{a \in C \mid c \leq a\}$
Adhikari et al. (2015)	$IC_{Adhikari}(c) = \frac{\log(depth(c)+1)}{\log(depth_{max}+1)} \times \left(1 - \log\left(\frac{\frac{ Leaves(c) \times nmih(c) }{Leaves_{max}}}{ subsumers'(c) } + 1\right)\right)$ $\times \left(1 - \frac{\log\left(1 + \sum_{a \in Hypo(c)} \frac{1}{depth(a)}\right)}{\log(Node_{max})}\right)$ $subsumers'(c) = subsumers(c) \cup \{c\}$

Table 4: State-of-the-art Information Content models evaluated in our experiments.

baseline, (Lastra-Díaz and García-Serrano, 2015a, fig.2).

In order to overcome the lexical coverage limitation associated to the ontologies, we argue that at least two strategies could be explored. The first strategy is the ontology population based on WordNet by using any automatic WordNet-based semantic annotation method, such as that explored by Sanfilippo et al. (2005). A second strategy is the automatic assembly of broad coverage “is-a” taxonomies from a large corpus such as Wikipedia, as is recently proposed and evaluated by Ben Aouicha et al. (2016a).

Finally, despite the plethora of ontology-based similarity measures and IC models available in the literature, the selection of a specific similarity measure for a particular application is still an open problem. For instance, a recent benchmark in a biomedical ontology-based IR task by Alonso and Contreras (2016) proves that there is no a statistically significant difference in performance between the *intrinsic IC* measure in (Garla and Brandt,

2012, eq. (13)) and the similarity measure introduced by Pedersen et al. (2007). This latter fact questions the extrapolation of the results and conclusions obtained in classic word similarity benchmarks to specific similarity-based applications. Thus, in order to improve our understanding of the problem, we suggest that the evaluation methodology of ontology-based similarity measures should be reconsidered by defining new task-oriented benchmarks and larger datasets. In this latter line of research, Jurgens et al. (2015) introduce a new similarity evaluation method called Cross-Level Semantic Similarity (CLSS), whose aim is to measure the contribution of the degree of similarity between small language units to the semantic similarity between larger linguistic units. Precisely, Pilehvar and Navigli (2015) propose an unified method to compute the semantic similarity between items from multiple linguistic levels. On the other hand, Saif et al. (2014) have carried out a study on the impact of the incompleteness of some linguistic resources

Well-founded IC models	Definition
CondProbHypo	$IC_{CPHypo}(c_i) = -\log_2(p_{Hypo}(c_i))$ $p_{Hypo}(c_i c_j) = \frac{1}{\sum_{\forall c_k \mid c_j \in LA(c_k)} (Hypo(c_k) +1)}$
CondProbUniform	$IC_{CPUni}(c_i) = -\log_2(p_{Uniform}(c_i))$ $p_{Uniform}(c_i c_j) = \frac{1}{ children(c_j) }$
CondProbLeaves	$IC_{CPLea}(c_i) = -\log_2(p_{Leaves}(c_i))$ $p_{Leaves}(c_i c_j) = \frac{1}{\sum_{\forall c_k \mid c_j \in LA(c_k)} (Leaves(c_k) +1)}$
CondProbLogistic	$IC_{CPLog}(c_i) = -\log_2(p_{Log}(c_i))$ $p_{Log}(c_i c_j) = \varphi_l(x) \circ p_{Hypo}(c_i c_j)$ $\varphi_l(x : k) = \frac{1}{1+e^{-k(x-\frac{1}{2})}}, \quad k^* = 8$
CondProbCosine	$IC_{CPCos}(c_i) = -\log_2(p_{Cos}(c_i))$ $p_{Cos}(c_i c_j) = \varphi_c(x) \circ p_{Hypo}(c_i c_j)$ $\varphi_c(x) = 1 - \cos\left(\frac{\pi}{2}x\right)$
CondProbCorpus	$IC_{CPCorpus}(c_i) = -\log_2(p(c_i))$ $p(c_i) = \begin{cases} 1 & , c_i = \Gamma \\ \sum_{\forall c_j \in LA(c_i)} p(c_j) p_{corpus}(c_i c_j) & , c_i \neq \Gamma \end{cases}$ $p_{corpus}(c_i c_j) = \frac{\max\{1, f(c_i)\}}{\sum_{\forall c_k \mid c_j \in LA(c_k)} \max\{1, f(c_k)\}}$

Table 5: Our current family of well-founded IC models introduced by [Lastra-Díaz and García-Serrano \(2015a\)](#) and evaluated in this work. $Hypo(c_i)$ and $Leaves(c_i)$ denote respectively the set of subsumed concepts and leaf concepts for any concept $c_i \in C$, without including the base concept c_i .

in Arabic, such as WordNet and Wikipedia, and into the performance of the ontology-based and gloss-based similarity measures. This latter work shows degradation of the performance from most ontology-based similarity measures, which call our attention to the problems of extrapolating the results based on English benchmarks and resources. Another interesting issue is the availability of a large word similarity benchmark based on WordNet that would also include instances of concepts and multiple-word terms, in the spirit of the TR9856 dataset introduced by [Levy et al. \(2015\)](#).

In summary, the mainstream of research into ontology-based similarity measures is still the proposal of new intrinsic IC models and IC-based measures, such as that proposed by [Pirró and Euzenat \(2010\)](#), [Meng et al. \(2014\)](#), [Gao et al. \(2015\)](#) and our aforementioned works. However, we also find in the literature some new corpus-based IC models such as that introduced by [Harispe et al. \(2015a\)](#), and some relevant non IC-based measures such as that proposed by [Sánchez et al. \(2012\)](#) and [Hadj Taieb et al. \(2014b\)](#). In addition, there are several strategies that could be explored in order to overcome the lexical coverage limitation of the ontologies, and the selection of a specific similarity measure for a particular application is still an open problem.

3.1 Related work on IC models

In another recent work by [Lastra-Díaz and García-Serrano \(2015a\)](#), we provide an in-depth review of the state of the art in IC models. For this reason, this section only provides a summary of the literature on IC models, including a review of the latest IC models published after our aforementioned work.

In [Resnik \(1995\)](#) and subsequently [Resnik \(1999\)](#), the

author introduces the first IC model reported in the literature. The Resnik IC model relies on a frequency counting method of the occurrences of a concept and its subsumed concepts into a corpus, that is also described in detail by [Pedersen, 2013, p.34](#), who uses the Resnik method to build the WordNet-based frequency files used in our experiments, [Pedersen \(2008b\)](#). The Resnik frequency counting method does not take the word senses into account; however, [Pedersen \(2010\)](#) proves that the IC models derived from a non sense-tagged corpus perform better than the sense-tagged ones. In order to overcome the drawbacks of the corpus-based IC models, [Seco et al. \(2004\)](#) introduce the first intrinsic IC model reported in the literature, whose core idea is that the IC models can be computed using only taxonomical features, such as the hyponym set ratio. During the last decade, the development of intrinsic IC models has become one of the mainstreams of research in the area. Among the main intrinsic IC models proposed in the literature, we find the works in [Zhou et al. \(2008a\)](#), [Sebti and Barfroush \(2008\)](#), [Blanchard et al. \(2008\)](#), [Sánchez et al. \(2011\)](#), [Sánchez and Batet \(2012\)](#), [Yuan et al. \(2013\)](#), and [Hadj Taieb et al. \(2014a\)](#), as shown in table 4, as well as the IC models introduced by [Lastra-Díaz and García-Serrano \(2015a\)](#) that are shown in table 5.

Finally, we have four recent works on IC models introduced by [Adhikari et al. \(2015\)](#), [Harispe et al. \(2015a\)](#), [Aouicha and Taieb \(2015\)](#) and [Ben Aouicha et al. \(2016b\)](#). First, [Harispe et al. \(2015a\)](#) introduce a family of corpus-based IC models based on the Belief function theoretical framework which is encouraged by the observation that the occurrences of a concept not only impact the IC value of the more general ancestor concepts, the so-called ancestors, but should also im-

pact the IC value of the more specific concepts, the so-called descendants. [Harispe et al. \(2015a\)](#) propose three different corpus-based IC models based on an adaptation of the classic *belief* and *plausibility* functions in the Demster-Shafer theory (DST), and the *pignistic* function. Second, [Adhikari et al. \(2015\)](#) introduce a new intrinsic IC model which is encouraged by the lack of integration in the previous IC models of a large combination of taxonomical features in order to distinguish several structural differences between concepts not considered before. The [Adhikari et al. \(2015\)](#) IC model integrates the relative depth, hyponym structure, subsumed leaves count and subsumer set count. [Aouicha and Taieb \(2015\)](#) introduce a new intrinsic IC model specifically designed for the MeSH biomedical ontology which has not been evaluated in WordNet. And finally, [Ben Aouicha et al. \(2016b\)](#) introduce a new intrinsic IC model on WordNet which is based on a new quantification of the ancestor set of each base concept. has not been included in our experiments. Tables 4 and 5 show the set of IC models that is implemented and evaluated in our experiments. This latest set of IC models, together with the recent IC models proposed by [Harispe et al. \(2015a\)](#) and [Aouicha and Taieb \(2015\)](#), represent, to the best of our knowledge, all the intrinsic and corpus-based IC models reported in the literature. On the other hand, [Blanchard et al. \(2008\)](#) IC_g is evaluated herein for the first time in a word similarity benchmark.

4 The proposed refinement

In [Lastra-Díaz and García-Serrano \(2015a\)](#), we propose a general framework to design IC models based on different methods for the estimation of the conditional probability between child and parent concepts, and we introduce a new family of IC models based on it, the so-called *well-founded IC models* shown in table 5. Our *IC models* are computed into four steps: (a) estimation of the conditional probabilities $p(c_i|c_j)$; (b) building of a total ordering of the concept set; (c) recovery of the concept probabilities $p(c_i)$ by using the recursive formula in equation (3); and (d) recovery of the IC values from the concept probabilities $p(c_i)$.

In order to eliminate the two drawbacks detailed in section 1.2, we introduce two refinements into the family of well-founded IC models and derive nine new IC models. First, in order to solve the problem related to the two cognitive IC models, we define a subsequent *normalization step* in the recovery of the concept probabilities in step (c) above, such that the overall sum of the probability on the leaf concepts is always 1 for these cases. Second, in order to warrant that the IC models satisfy the *growing monotonicity axiom*, such that $\forall c_i \subseteq c_j \Rightarrow IC(c_i) \geq IC(c_j)$, we define a new method for recovering the final concept probabilities based on the definition of the probability $p(c_i)$ as the sum of the probabilities of the leaf concepts subsumed by the concept c_i , instead of the direct value returned by the recursive formula in equation (3). Thus, we define a subsequent *subsumed probability recovery* step in the probability recovery step (d) above. We note that this new definition

of the concept probabilities as the probability of their subsumed leaves matches the axiomatic construction of a discrete probability space, as introduced by [Lastra-Díaz and García-Serrano \(2015a\)](#), or any book on the subject, such as [Ash and Doléans-Dade \(2000\)](#). The new method to compute the final probabilities $p(c_i)$ from the conditional probabilities $p(c_i|c_j)$ matches the previous method in our aforementioned work whenever the taxonomy is tree-like, but it produces a slightly different probability function on taxonomies with multiple inheritance. This latest refinement is a sufficient condition to satisfy the *growing monotonicity axiom* regardless of the conditional probability model or the type of base taxonomy.

Refinement 1. In order to satisfy the growing monotonicity axiom regardless of the type of taxonomy, we introduce the following changes into the algorithm used to build the well-founded IC models. First, we introduce the growing monotonicity axiom as a further axiom into the definition of a *well-founded IC model*. And second, in order to satisfy the new axiom (4) the concept probability is defined as the sum of the probability of its subsumed leaves, instead of the direct value obtained from the recursive formula in equation (3), as was done in our aforementioned work.

Refinement 2. In order to warrant that the sum of leaf concept probabilities is 1 for any cognitive IC model, such as the *CondProbLogistic* and *CondProbCosine* introduced in [Lastra-Díaz and García-Serrano \(2015a\)](#), it is necessary to normalize the overall sum of leaf probabilities to 1.

All new IC models share the same algebraic and computational structure, being computed into six steps: (1) estimation of the conditional probabilities; (2) building of a total ordering of the concepts within the taxonomy; (3) recovery of the concept probabilities $p(c_i)$ by using the recursive formula in equation (3); (4) *unit normalization* of the probability of the leaf nodes only for the IC models based on non-linear transformations of the conditional probability; (5) computation of each concept probability $p(c_i)$ as the overall sum of the probability of its subsumed leaves; and finally, (6) computation of the IC values from the concept probabilities. In this way, the new steps (4) and (5) above eliminate the two aforementioned drawbacks, but the four remaining steps are identical to the original algorithm 1 in our previous work.

The two refinements above lead us to the reformulation of the *algorithm 1* to build the well-founded IC models introduced by [Lastra-Díaz and García-Serrano \(2015a\)](#). The previous *algorithm 1* is substituted by the *new algorithm* to build the well-founded IC models, which is summarized in table 6. Unlike the previous *algorithm 1*, the *new algorithm* only uses the iterative top-down procedure defined by the recursive formula in equation (3) in order to compute the probability of the leaf nodes, not the probability of each concept as was done in our aforementioned work. We recall that the probability recovery algorithm defined by the top-down

formula in equation (3) warrants that the overall sum of the leaf probabilities is 1 if the conditional probabilities $p(c_i|c_j)$ are well-defined and satisfy the constraint in equation (1). This latter fact is formalized into the proposition 2 below.

The *New Algorithm* in table 6 works on any type of taxonomy, and satisfies all the structure axioms in definition 1. The algorithm includes the two modifications proposed above in order to eliminate the two drawbacks found in our previous method. Thus, the proposed algorithm completely closes the algebraic and computational definition of the family of well-founded IC models, and it should be used in the design of any new intrinsic IC model.

Definition 1 (refined well-founded IC model)

Given a taxonomy of concepts $\mathcal{C} = (C, \leq_C, \Gamma)$, and an IC model defined by the function $IC : C \rightarrow \mathbb{R}^+ \cup \{0\}$, we call it a refined well-founded IC model if it can be written as $IC(c) = -\log_2(p(c))$ where $p(c)$ is a concept-valued function as defined in equation (4), and the functions $p(c_i|c_j)$ are the conditional probabilities between any child concept c_i and its parent concepts c_j , which satisfy the edge-based property as defined in equation (1).

- (1) *Edge-based axiom.* The sum of conditional probabilities $p(c_i|c_j)$ of the children nodes c_i on any parent c_j node must be equal to 1, as defined in equation (1), where $LA(c_i)$ denotes the set of lowest ancestors (direct parents) of any concept c_i .

$$\sum_{\forall c_i|c_j \in LA(c_i)} p(c_i|c_j) = 1 \tag{1}$$

- (2) *Leaf node probability axiom.* The overall probability of the leaf concepts sums 1, as defined in equation (2), and they are computed using the iterative top-down algorithm defined by equation (3).

$$\sum_{c_k \in L_C} p(c_k) = 1 \tag{2}$$

$$p : C \rightarrow [0, 1] \subset \mathbb{R}$$

$$p(c_i) = \begin{cases} 1 & , c_i = \Gamma \\ \sum_{\forall c_j \in LA(c_i)} p(c_j)p(c_i|c_j) & , c_i \neq \Gamma \end{cases} \tag{3}$$

- (3) *Probability node axiom.* The probability $p(c_i)$ for each concept $c_i \in C$ must be equal to the sum of the probability of each sub-summed leaf concept $c_k \in Leaves(c_i) = \{c_k \in C \mid c_k \leq_C c_i \wedge c_k \text{ is a leaf concept}\}$, as defined in equation (4).

$$p(c_i) = \sum_{c_k \in Leaves(c_i)} p(c_k) \tag{4}$$

- (4) *Monotonicity.* $\forall c_i, c_j \in C, c_i \leq_C c_j \Rightarrow IC(c_i) \geq IC(c_j)$

The axioms (1), (2) and (3) above allow us to define a new family of well-founded intrinsic IC models based on the estimation of the conditional probabilities $p(c_i|c_j)$ for each edge of the taxonomy, as shown in table 7. The axiom (3) is a sufficient condition for the satisfaction of the axiom (4), thus, the new refined IC models satisfy the monotonicity axiom by design. We call the new family as *refined well-founded IC models* in order to distinguish it from our previous IC models, and to emphasize the use of the new algorithm in table 6. In proposition 1, we show that given a taxonomy (C, \leq_C, Γ) , the definition of the concept probabilities according to axiom (3) is a sufficient condition to get a well-founded probability space, which moreover matches the standard axiomatic construction of any discrete probability space. In addition, we show in proposition 2 that axioms (1) and (2) of a well-founded IC model are sufficient conditions to build a leaf-valued function $p : L_C \subset C \rightarrow [0, 1]$ that satisfies axiom (2) above and the second premise of proposition 1. Thus, proposition 2 proves that any well-founded IC model induces a well-founded probability space on any base taxonomy, and the whole system is supported by the structures derived from the conditional probabilities. The proofs of both propositions are included in appendix B of Lastra-Díaz and García-Serrano (2015a).

Proposition 1 *Be a taxonomy $\mathcal{C} = (C, \leq_C, \Gamma)$ defined by a partially ordered set (C, \leq_C) with a distinguished supreme element Γ , called the root, and L_C the set of leaves in C . If a set-valued positive function P is defined from the leaf-valued function p as follows:*

$$P : 2^\Gamma \rightarrow [0, 1]$$

$$(1) \quad P(A) = \sum_{c_k \in L_C \cap A} p(c_k)$$

$$p : L_C \subset C \rightarrow [0, 1]$$

$$(2) \quad \sum_{c_k \in L_C} p(c_k) = 1$$

then the following facts are satisfied: (1) P is a probability measure, and (2) the triplet $(\Gamma, 2^\Gamma, P)$ is a probability space.

Proposition 2 *Let a taxonomy $\mathcal{C} = (C, \leq_C, \Gamma)$ and L_C be the set of leaves in C . Given a concept-valued function p defined by*

$$p : C \rightarrow [0, 1]$$

$$p(c_i) = \begin{cases} 1 & , \text{if } c_i = \Gamma \\ \sum_{\forall c_j \in LA_C(c_i)} p(c_i|c_j)p(c_j) & , \text{otherwise} \end{cases}$$

then $P(L_C) = 1$, as given below:

$$P(L_C) = \sum_{c_k \in L_C} p(c_k) = 1$$

4.1 The new family of IC models

This section introduces eight new intrinsic IC models called *CondProbRefHyponyms*, *CondProbRefUniform*, *CondProbRefLeaves*, *CondProbRefLogistic*,

New probability and IC recovery algorithm	
Input:	a rooted taxonomy $C = (C, \leq_C, \Gamma)$ (1) $p(c_i c_j)$ for each child and parent concepts.
Output:	(2) $p : C \rightarrow [0, 1] \subset \mathbb{R}$ (3) $IC : C \times C \rightarrow \mathbb{R}^+ \cup \{0\}$
1:	Compute the conditional probabilities $p(c_i c_j)$.
2:	Build a queue Q with a total ordering of the taxonomy (C, \leq_C, Γ) , such that every concept is in a subsequent position to every one of its parent concepts <i>Remark:</i> top-down computation of the leaf node probabilities
3:	foreach $c_i \in Q$
4:	$p(c_i) = \begin{cases} 1 & , \text{ if } c_i = \Gamma \\ \sum_{\forall c_j \in LA_C(c_i)} p(c_i c_j) p(c_j) & , \text{ otherwise} \end{cases}$
5:	end foreach <i>Remark:</i> normalization of the overall leaf node probability (only if the $p(c_i c_j)$ values do not satisfy axiom 1)
6:	$overallLeavesProb = \sum_{c_k \in Leaves(\Gamma)} p(c_k)$
7:	foreach $c_i \in Leaves(\Gamma)$
8:	$p(c_i) = \frac{p(c_i)}{overallLeavesProb}$
9:	end foreach <i>Remark:</i> bottom-up computation of the node probabilities <i>Remark:</i> for the computation of the probability of each node, $Leaves(c_i)$ denotes the set of subsumed leaf concepts inclusive c_i .
10:	foreach $c_i \in Q$
11:	$p(c_i) = \sum_{c_k \in Leaves(c_i)} p(c_k)$
12:	$IC(c_i) = -\log_2 p(c_i)$
13:	end foreach

Table 6: New algorithm for the computation of the refined well-founded IC models.

CondProbRefCosine, *CondProbLogisticLeaves*, *CondProbRefCosineLeaves* and *CondProbRefLeavesSubsumersRatio*, and a new corpus-based IC model called *CondProbRefCorpus*. From the latter list, the first five intrinsic IC models and the *CondProbRefCorpus* IC model are derived from the corresponding IC models introduced by [Lastra-Díaz and García-Serrano \(2015a\)](#) by using the *new algorithm* to compute the probability and IC values detailed in table 6. On the other hand, the new intrinsic IC models called *CondProbLogisticLeaves*, *CondProbRefCosineLeaves* and *CondProbRefLeavesSubsumersRatio* are based on three new methods to estimate the conditional probabilities $p(c_i|c_j)$. The *CondProbLogisticLeaves* and *CondProbRefCosineLeaves* IC models combine the conditional probability function $p_{Leaves}(c_i|c_j)$ with two different cognitive-based non-linear similarity functions previously introduced in our aforementioned work.

Because of the good performance exhibited by the [Sánchez et al. \(2011\)](#) IC model in combination with our *coswJ&C* similarity measure, we propose the *CondProbRefLeavesSubsumersRatio* IC model which is a reformulation of the [Sánchez et al. \(2011\)](#) IC model based on the general framework proposed by the family of IC models introduced herein. This new IC model is based on the fact that the difference in IC values between child and parent concepts in a tree-like taxonomy matches the

logarithm of the conditional probability $p(c_i|c_j)$. This latest observation inspired the family of IC-based similarity measures introduced by [Lastra-Díaz and García-Serrano \(2015b\)](#), and from it follows that the [Sánchez et al. \(2011\)](#) IC model can be reformulated as the ratio between child and parent concepts of the function $\sigma(x)$ in table 7. The function $\sigma(x)$ is called *Sánchez-Batet-Isern estimator*, because $\sigma(x)$ can be interpreted as a taxonomical estimator of the concept probabilities. Precisely, the *CondProbRefLeavesSubsumersRatio* IC model defines a well-defined probability space from the kernel function of the [Sánchez et al. \(2011\)](#) IC model, and this same strategy could be used in order to reformulate other IC models, or taxonomy-based conditional probability estimators, in the general framework proposed by our family of IC models.

Table 7 shows the definition of the new family of IC models. For the formulas in table 7, $Hypo(c_i)$ and $Leaves(c_i)$ denote respectively the set of subsumed concepts and subsumed leaf concepts for any concept $c_i \in C$, without including the base concept c_i . Unlike our previous work, each concept probability denoted by $p^*(c_i)$ is defined as the sum of the probability of the subsumed leaf nodes in equation (4), instead of the value directly obtained from the top-down formula in equation (3). The probability values $p(c_i)$ of the non-leaf concepts that are obtained from the top-down formula in equa-

New IC models in this work	Definition
CondProbRefHyponym	$IC_{CPRefHypo}(c_i) = -\log_2(p_{Hypo}^*(c_i))$ $p_{Hypo}(c_i c_j) = \frac{ Hypo(c_i) +1}{\sum_{\forall c_k c_j \in LA(c_k)} (Hypo(c_k) +1)}$
CondProbRefUniform	$IC_{CPRefUni}(c_i) = -\log_2(p_{Uniform}^*(c_i))$ $p_{Uniform}(c_i c_j) = \frac{1}{ children(c_j) }$
CondProbRefLeaves	$IC_{CPRefLea}(c_i) = -\log_2(p_{Leaves}^*(c_i))$ $p_{Leaves}(c_i c_j) = \frac{ Leaves(c_i) +1}{\sum_{\forall c_k c_j \in LA(c_k)} (Leaves(c_k) +1)}$
CondProbRefLogistic	$IC_{CPRefLog}(c_i) = -\log_2(p_{Log}^*(c_i))$ $p_{Log}(c_i c_j) = \varphi_l(x) \circ p_{Hypo}(c_i c_j)$ $\varphi_l(x : k) = \frac{1}{1+e^{-k(x-\frac{1}{2})}}, \quad k^* = 8$
CondProbRefCosine	$IC_{CPRefCos}(c_i) = -\log_2(p_{Cos}^*(c_i))$ $p_{Cos}(c_i c_j) = \varphi_c(x) \circ p_{Hypo}(c_i c_j)$ $\varphi_c(x) = 1 - \cos\left(\frac{\pi}{2}x\right)$
CondProbRefCorpus	$IC_{CPRefCorpus}(c_i) = -\log_2(p^*(c_i))$ $p_{corpus}(c_i c_j) = \frac{\max\{1, f(c_i)\}}{\sum_{\forall c_k c_j \in LA(c_k)} \max\{1, f(c_k)\}}$
CondProbRefLogisticLeaves	$IC_{CPRefLogLeaves}(c_i) = -\log_2(p_{LogLeaves}^*(c_i))$ $p_{LogLeaves}(c_i c_j) = \varphi_l(x) \circ p_{Leaves}(c_i c_j)$ $\varphi_l(x : k) = \frac{1}{1+e^{-k(x-\frac{1}{2})}}, \quad k^* = 8$
CondProbRefCosineLeaves	$IC_{CPRefCosLeaves}(c_i) = -\log_2(p_{CosLeaves}^*(c_i))$ $p_{CosLeaves}(c_i c_j) = \varphi_c(x) \circ p_{Leaves}(c_i c_j)$ $\varphi_c(x) = 1 - \cos\left(\frac{\pi}{2}x\right)$
CondProbRefLeavesSubsumersRatio	$IC_{CPRefLeaSubRat}(c_i) = -\log_2(p_{LeaSubRat}^*(c_i))$ $p_{LeaSubRat}(c_i c_j) = \frac{\frac{\sigma(c_i)}{\sigma(c_j)}}{\sum_{\forall c_k c_j \in LA(c_k)} \frac{\sigma(c_k)}{\sigma(c_j)}}$ $\sigma(c) = \frac{ Leaves(c) }{ subsumers(c) } + 1$

Table 7: New set of IC models proposed into the family of well-founded IC models. Unlike our previous work, each concept probability denoted by $p^*(c_i)$ is defined as the sum of the probability of the subsumed leaf nodes, instead of the value directly obtained from the recursive formula in equation (3). The new IC models are computed using the new algorithm detailed in Table 5. $Hypo(c_i)$ and $Leaves(c_i)$ denote respectively the set of subsumed concepts and leaf concepts for any concept $c_i \in C$, without including the base concept c_i .

tion (3) are only temporary values whose aim is to obtain the estimated probability value of each leaf concept. The new IC models are computed using the new algorithm detailed in table 6. The *CondProbRefLogistic*, *CondProbRefCosine*, *CondProbLogisticLeaves* and *CondProbRefCosineLeaves* IC models do not satisfy the edge-based axiom defined by equation (1) in definition 1 because of they integrate a non-linear monotone transformation in their definition that prevents it, thus, the weights of the taxonomy used with the *coswJ&C* similarity measure in table 2 are set to $|IC(c_i) - IC(c_j)|$ instead of $-\log_2(p(c_i|c_j))$.

5 Evaluation

The goals of the experiments described in this section are as follows: (1) the experimental evaluation of the proposed IC models and their comparison with the state-of-the-art methods; (2) a new experimental study onto the state of the art in ontology-based similarity measures; (3) a detailed statistical significance analysis of the similarity measures and IC models; (4) the replica-

tion of previously reported methods and results; (5) a new comparison between intrinsic and corpus-based IC models; (6) a study into the impact of the IC models on the IC-based similarity measures; (7) a comparison of the computational cost of the ontology-based similarity measures; (8) a new confirmation of the findings in our previous aforementioned works on the refuted outperformance of the intrinsic IC models over the corpus-based ones; and (9) a new confirmation of the achievements of the family of intrinsic IC models and IC-based similarity measures.

5.1 Methods evaluated herein

In order to compare the new family of IC models in table 7 with the state-of-the-art IC models, as well as providing a conclusive image of the state of the art of the problem, we implemented and evaluated all the IC models in tables 4, 5 and 7, as well as all the IC-based similarity measures in table 2 and the remaining ontology-based similarity measures shown in table 1. One IC model introduced by Blanchard et al. (2008) is evaluated herein for the first time. To the best of our knowledge, we

evaluate herein all WordNet-based intrinsic IC models reported in the literature, with the only exception of the IC model very recently introduced by [Ben Aouicha et al. \(2016b\)](#). Therefore, the experiments reported herein are the largest experimental survey of intrinsic IC models and ontology-based similarity measures reported up to date, which are based on a same code implementation.

For all the similarity measures and IC models, the depth is defined as the length of the shortest ascending path from each concept to the root. For the Zhou et al. IC model, the authors define the depth starting at 1 for the root concept. All methods have been implemented in a Java software library called HESML, which has been developed by the authors in order to replicate all methods evaluated herein. HESML was also used in our two aforementioned works on IC-based similarity measures and IC models, and it is going to be introduced and released in another forthcoming paper, [Lastra-Díaz and García-Serrano \(2016\)](#), together with a set of reproducible experiments and a replication dataset called *WNSimRep v1*.

In order to compare the intrinsic and corpus-based IC models, we use as baseline a corpus-based Resnik IC model based on the Wordnet-based frequency file called “ic-treebank-add1.dat” included in [Pedersen \(2008b\)](#), which was also used as a baseline in [Lastra-Díaz and García-Serrano \(2015a\)](#), having been the best performing corpus-based IC model in [Lastra-Díaz and García-Serrano \(2015b\)](#).

5.2 Experimental setup

We follow the same experimental setup defined by [Lastra-Díaz and García-Serrano \(2015a\)](#), including the same preprocessing steps, evaluation metrics, baselines, management of polysemic words and reporting of the results. In addition, we include for the first time a detailed pairwise statistical significance analysis between each pair of IC models and IC-based measures. We use the noun database of Wordnet 3.0, [Miller \(1995\)](#), and the five most significant word similarity benchmarks shown in table 8. For each word pair, we select the highest similarity value between the pairwise comparison of the sets of concepts evoked by each word.

Some preprocessing was necessary for the Agirre203 and SimLex-999 datasets to carry out the experiments. For the Agirre203 dataset, it was necessary to remove two word pairs containing verbs not present in the noun database of Wordnet 3.0, such as the pairs (*drink, eat*) and (*stock, live*). In addition, it was also necessary to change the term “media” for “medium”, and “children” for “child”, because these terms do not appear directly in noun database. For this reason, we only used 201 nouns instead of 203, thus, this subset is called hereafter Agirre201. In the case of SimLex-999, it contains 666 nouns, but the word “august” is not included as synset in WordNet 3.0, thus, we only used 665 nouns from the SimLex-999 dataset, and this subset is called hereafter SimLex665. Finally, the MC30 dataset introduced by [Miller and Charles \(1991\)](#) is made up by 30 noun pairs; however, two word pairs are commonly

Reference	Acronym	#wp	Description
Rubenstein and Goodenough (1965)	RG65	65	65 noun pairs ranging a similarity between 0 and 4.
Miller and Charles (1991)	MC28	28	Subset of RG65
Agirre et al. (2009)	Agirre201	201	Pure similarity subset of Finkelstein et al. (2002) with similarity in the range 0 to 10.
Pirr6 (2009)	$P\&S_{full}$	65	Modern replication of RG65
Hill et al. (2015)	SimLex665	665	Noun subset of SimLex-999 with similarity in the range 0 to 10.

Table 8: Word similarity benchmarks used in our experiments

omitted because of they were not included in previous versions of WordNet. For this reason, we use the MC28 dataset as defined at ([Resnik, 1995](#), table 3) and ([Li et al., 2003](#), p.875), together with the original human similarity judgements introduced by [Rubenstein and Goodenough \(1965\)](#). The datasets corresponding to the similarity benchmarks shown in table 8 are included in the HESML distribution.

5.3 Evaluation metrics

As evaluation metrics, we use the Pearson correlation factor, denoted by r in equation (5), and the Spearman rank correlation factor, denoted by ρ in equation (6). For a detailed review of the latter metrics, we refer the reader to ([Lastra-Díaz and García-Serrano, 2015a](#), §5.3).

$$r = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}} \quad (5)$$

$$\rho = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)}, \quad d_i = (x_i - y_i) \quad (6)$$

In order to compare the performance of the IC models, we use the average Pearson and Spearman correlation values for each pair (IC model, IC-based similarity measure) on all datasets. The statistical significance of the results is evaluated by using the p-values resulting from the t-student test for the difference mean between the Spearman correlation values reported by each pair of IC models or IC-based similarity measures. The p-values are computed by using a one-sided t-student distribution on two paired sample sets. For the p-values between IC models, we use the vectors of the average Spearman correlation values over each IC-based similarity measure (rows in table 11) as a paired sample set, whilst for the similarity measures we use the vectors of Spearman correlation values of each similarity measure over all datasets (rows in table 12). Our null hypothesis,

denoted by H_0 , is that the difference in the average performance between the compared IC models or IC-based measures is 0, whilst the alternative hypothesis, denoted by H_1 , is that their average performance is different. For a 5% level of significance, it means that if the p-value is greater than 0.05, we must accept the null hypothesis, otherwise we can reject H_0 with an error probability of less than the p-value.

The Spearman rank correlation metric can represent better the use of the similarity measures in most rank-based selection tasks in NLP and IR, because it “provides an evaluation metric that is independent of these data-dependent transformations”, (Agirre et al., 2009, §6). In addition, most similarity measures are monotone transformations from previous measures. Therefore, a statistical significance analysis based on the Spearman correlation shows the intrinsic differences and similarities between methods in a more conclusive manner than an analysis based on the Pearson correlation. Likewise, in order to compare the IC-based similarity measures, we selected for each measure its best performing IC model according to the average Spearman correlation values shown in table 11.

5.4 Results obtained

Table 9 below shows the computational cost of each similarity measure on the MC28 dataset. The remaining data tables are included in the appendix next to the bibliography. Tables 10 and 11 show in each cell the average Pearson and Spearman correlation values respectively obtained in the evaluation of each IC model with any IC-based similarity measure on all datasets. Table 12 shows the Pearson and Spearman correlation values obtained by each ontology-based similarity measure on all datasets. In order to make the interpretation of the resulting p-values easier, tables 13 and 14 show a summary of the statistical significance analysis between the IC models and ontology-based similarity measures, whilst the raw p-values are shown in tables 25 and 26. Each row in tables 13 and 14 shows an ‘x’ whenever the method in the row header obtains a statistically significant higher performance than the method in the column header. Thus, the rows show the methods that are outperformed by each method on the left, whilst the columns show the methods that outperform each method at the top. Finally, tables 15 to 24 in the appendix show all raw data tables for the cross-evaluation of the IC models and IC-based similarity measures on all datasets.

6 Discussion

6.1 Comparison of the IC models

Looking at tables 10 and 11, the following conclusions can be drawn: (1) the Seco et al. (2004) IC model obtains the highest average Pearson and Spearman correlation values on all datasets and IC-based similarity measures, as it is the best performing IC model on average; (2) a large set of IC models made up of the models introduced by Seco et al. (2004), Blanchard et al. (2008),

Similarity measure	Overall (msec)	Avg (msec)	Ratio
Sánchez et al. (2012)	480	17.14	0.66
Pirró and Seco (2008)	696	24.86	0.96
Pirró and Euzenat (2010)	703	25.11	0.97
Garla and Brandt (2012)	715	25.54	0.98
Meng and Gu (2012)	716	25.57	0.99
Jiang and Conrath (1997)	722	25.79	0.99
Resnik (1995) (baseline)	726	25.93	1.00
Lin (1998)	728	26.00	1.00
Lastra-Díaz and García-Serrano (2015b), cosJ&C	735	26.25	1.01
Hadj Taieb et al. (2014b)	774	27.64	1.07
Al-Mubaid and Nguyen (2009)	38016	1357.71	52.36
Wu and Palmer (1994)	42514	1518.36	58.56
Gao et al. (2015)	44343	1583.68	61.08
Li et al. (2003), strategy 9	45201	1614.32	62.26
Meng et al. (2014)	48499	1732.11	66.80
Zhou et al. (2008b)	50343	1797.96	69.34
Pedersen et al. (2007)	53504	1910.86	73.70
Leacock and Chodorow (1998)	53921	1925.75	74.27
Li et al. (2003), strategy 4	54278	1938.50	74.76
Li et al. (2003), strategy 3	54607	1950.25	75.22
Rada et al. (1989)	56172	2006.14	77.37
Lastra-Díaz and García-Serrano (2015b), coswJ&C	172490	6160.36	237.59

Table 9: Overall running time and average time per word pair for each similarity measure in the MC28 dataset with the following PC setup: Windows 8.1 x64, Java 1.8, Intel Core i7-5570 @ 2.40 GHz, 8 Gb RAM. The rows are arranged in ascending order according to the running time reported for each similarity measure. All the similarity measures have been implemented and evaluated within a same software library developed by the authors. The last row shows the running time ratio as regard the baseline defined by the Resnik measure.

Sánchez et al. (2011), Sánchez et al. (2012), Meng et al. (2012), Yuan et al. (2013) and Adhikari et al. (2015) obtain on average a higher Pearson and Spearman correlation values than the corpus-based IC model defined as baseline; (3) the new IC models called *CondProbRefHyponyms* and *CondProbRefCosine* obtain on average a higher Pearson and Spearman correlation values respectively than the baseline IC model, and the Zhou et al. (2008a) IC model also obtains on average a higher Spearman correlation value than the baseline IC model; (4) most of our family of well-founded IC models and the Hadj Taieb et al. (2014a) IC model obtain on average a lower Pearson and Spearman correlation values than the baseline IC model; and (5) the Hadj Taieb et al. (2014a) IC model obtains on average the lowest Pearson and Spearman correlation values among all IC models, and its average performance is much lower than the remaining IC models.

Tables 15 to 24 allow the following conclusions to be drawn: (1) the Sánchez et al. (2011) IC model obtains the highest Pearson correlation value with our *coswJ&C*

similarity measure in the RG65 dataset; (2) the Resnik IC model obtains the highest Pearson correlation value with the J&C similarity measure in the MC28 dataset; (3) our new *CondProbRefUniform* IC model obtains the highest Pearson correlation value with the FaITH similarity measure in the Agirre201 dataset; (4) the Yuan et al. (2013) IC model obtains the highest Pearson correlation value with the FaITH measure in the $P\&S_{full}$ dataset; and (5) the Seco et al. (2004) IC model obtains the highest Pearson correlation value with the Zhou et al. (2008b) similarity measure in the SimLex665 dataset. In addition, an analysis of the raw Spearman correlation values on all datasets allows the following conclusions to be drawn: (6) the Meng et al. (2012) IC model obtains the highest Spearman correlation value with our *coswJ&C* measure in the RG65 dataset; (7) the Resnik IC model obtains the highest Spearman correlation value with our *coswJ&C* measure in the MC28 dataset; (8) our new *CondProbRefUniform* IC model obtains the highest Spearman correlation value with the Lin (1998), FaITH and Meng and Gu (2012) similarity measures in the Agirre201 dataset; (9) the Sánchez et al. (2011) IC model obtains the highest Spearman correlation value with our *coswJ&C* similarity measure in the $P\&S_{full}$ dataset; and (10) the Yuan et al. (2013) IC model obtains the highest Spearman correlation value with the Zhou et al. (2008b) similarity measure in the SimLex665 dataset.

6.2 The statistical significance of the IC models

Table 13 allows the following conclusions to be drawn. First, the Seco et al. (2004) IC model obtains a statistically significant higher average Spearman correlation value than the remaining IC models with the only exception of the Sánchez et al. (2011) IC model. Second, Seco et al. (2004) and Sánchez et al. (2011) are the only IC models that are not outperformed in a statistically significant manner by another IC model. Third, the Seco et al. (2004), Sánchez et al. (2011) and Yuan et al. (2013) IC models obtain a statistically significant higher average Spearman correlation value than the baseline defined by the corpus-based Resnik IC model, thus, this small set of state-of-the-art intrinsic IC models outperform the best performing corpus-based IC model, confirming the H3 hypothesis positively. Fourth, the Hadj Taieb et al. (2014a) IC model obtains a statistically significant lower average Spearman correlation than all of the IC models. Fifth, most of our intrinsic IC models obtain a statistically significant lower average Spearman correlation than the rest of the IC models, with the exception of the *CondProbHyponyms*, *CondProbCosine*, *CondProbRefHyponyms*, *CondProbRefLeaves* and *CondProbRefCosine* IC models. Sixth, the Zhou et al. (2008a), Meng et al. (2012) and Yuan et al. (2013) IC models only obtain a statistically significant lower average Spearman correlation than the Seco et al. (2004) IC model, whilst the Adhikari et al. (2015) IC model is only outperformed by another two IC models. Thus, the Zhou et al. (2008a), Meng et al. (2012), Yuan et al. (2013) and

Adhikari et al. (2015) IC models follow the Seco et al. (2004) and Sánchez et al. (2011) IC models in terms of performance in the average Spearman correlation. However, looking at table 10, we see that the performance measured by the Pearson correlation of the Zhou et al. (2008a) IC model is much lower than the remaining IC models. And seventh, among the twenty-five IC models analyzed, the Resnik IC model defined as baseline obtains a statistically significant higher average Spearman correlation than other ten models, and it is statistically outperformed by only three intrinsic IC models, thus, there is no a statistically significant difference between most intrinsic IC models and the baseline, a fact that confirms the hypothesis H2 positively.

Finally, the hypothesis H8 behind the refinement and the new IC models introduced in this work is positively confirmed by the data obtained in our experiments. Looking at table 13, we see that the new IC models *CondProbRefUniform*, *CondProbRefLeaves*, *CondProbRefCosine* and *CondProbRefCorpus*, obtain a statistically significant higher average Spearman correlation than their corresponding non-refined IC models *CondProbUniform*, *CondProbLeaves*, *CondProbCosine* and *CondProbCorpus*. However, the *CondProbRefHyponyms* and *CondProbRefLogistic* IC models are not able to obtain a statistically significant higher performance than their corresponding models *CondProbHyponyms* and *CondProbLogistic*.

6.3 Comparison of the similarity measures

Table 12 shows that our *coswJ&C* similarity measure combined with the Sánchez et al. (2011) IC model obtains the highest Spearman correlation values in all datasets, with the only exception of SimLex665, the highest Pearson correlation values in the RG65 (0.8870) and MC28 (0.8710) datasets, as well as the highest overall average combined Pearson and Spearman correlation values (0.7708) shown in the last column and the highest overall average Spearman correlation value (0.7579). We point out that the highest Pearson correlation value (0.8809) in the MC28 dataset is obtained by the J&C similarity measure with the Resnik IC model, as shown in table 17, whilst the Seco et al. (2004) IC model is used for the overall comparison in table 12, because this latter IC model is the best performing IC model for the J&C measure in terms of the Spearman correlation.

Table 12 also shows that the Zhou et al. (2008b) similarity measure obtains the highest Pearson (0.6237) and Spearman (0.6101) correlation values in the SimLex665 dataset and the highest overall average Pearson correlation value (0.7859). In addition, the Zhou et al. (2008b) measure obtains the second best overall performance. The Hadj Taieb et al. (2014b) similarity measure obtains the highest Pearson correlation value (0.7123) in the Agirre201 dataset. The FaITH similarity measure introduced by Pirró and Euzenat (2010) obtains the highest Pearson correlation value (0.9082) in the $P\&S_{full}$ dataset when it is combined with the Yuan et al. (2013) model.

Table 12 shows that a small set of similarity measures obtain a higher overall performance than the baseline defined by J&C measure, as well as the Resnik and Lin similarity measures. This small set of outperforming measures is made up of our *coswJ&C* and *cosJ&C* similarity measures and the measures introduced by Zhou et al. (2008b), Pirró and Seco (2008), Hadj Taieb et al. (2014b) and Gao et al. (2015). In addition, a large set of ontology-based similarity measures obtain a higher average Pearson correlation value than the baseline defined by the J&C similarity measure.

The *coswJ&C* similarity measure, in combination with the Sánchez et al. (2011) IC model, obtains the best overall performance defined by the average of the Pearson and Spearman correlation values, as shown in last column of table 12. In addition, the *coswJ&C* similarity measure outperforms the remaining measures in the Spearman correlation metric. Looking at table 18, we can see another very meaningful and unexpected fact: *the coswJ&C similarity measure obtains the highest Spearman correlation value in the MC28 dataset with all the IC models, excluding the Hadj Taieb et al. (2014a) IC model.* We attribute the good performance of the *coswJ&C* similarity measure in the Spearman metric to the novel method for computing the distance between concepts that is defined by our distance $dis_{wJ&C}$ introduced in Lastra-Díaz (2014) and Lastra-Díaz and García-Serrano (2015b), which defines an IC-based weighted graph as a generalization of the classic Jiang-Conrath distance. On the other hand, this $dis_{wJ&C}$ measure requires the computation of the length of the shortest path on a non-uniform and real-valued weighted graph using the Dijkstra algorithm, whose computation time is longer than for the case in which only the edge count is required, as happens for the rest of the hybrid IC-based similarity measures shown in table 2. For this reason, the *coswJ&C* measure reports the highest computational cost in table 9, which is roughly three times greater than most hybrid IC-based similarity measures.

The data in table 9 allows the hypothesis H5 and the following conclusion introduced in Lastra-Díaz and García-Serrano (2015b) to be confirmed: despite the *coswJ&C* and Zhou et al. (2008b) similarity measures outperforming the remaining similarity measures on average, the computational cost and the performance of these measures, as well as the remaining hybrid IC-based similarity measures, prevent their use in practical applications. Thus, a practical option is to use our *cosJ&C* similarity measure, which obtains the third best overall performance, despite there being no statistical significant difference between it and the measures introduced by Pirró and Seco (2008) and Hadj Taieb et al. (2014b). Indeed, the general conclusion that we advance here is that the performance margin between the state-of-the-art ontology-based similarity measures is very narrow.

An interesting point is that the three similarity measures on top of table 12 are derived from the Jiang-Conrath distance. The *coswJ&C* similarity measure is a generalization of the Jiang-Conrath measure based on an IC-based weighted graph, whilst the Zhou et al.

(2008b) similarity measure is a linear combination of it with the Leacock and Chodorow (1998) similarity measure. On the other hand, the *cosJ&C* similarity measure is a monotone transformation of the Jiang-Conrath distance. Thus, the measurement strategy introduced by Jiang and Conrath (1997) leads the state of the art of the problem.

6.4 The statistical significance of the similarity measures

Table 14 allows the following conclusions to be drawn: (1), our *coswJ&C* similarity measure and the measure introduced by Zhou et al. (2008b) obtain a statistically significant higher average Spearman correlation value than the baseline defined by the J&C measure, and they are the only measures that outperform the baseline; (2) our *coswJ&C* similarity measure, and the measures introduced by Zhou et al. (2008b) and Meng et al. (2014), are the only measures that are not outperformed by other measures in a statistically significant manner; (3) the Zhou et al. (2008b) similarity measure obtains a statistically significant higher average Spearman correlation value than all of the measures, with the only exception of the *coswJ&C* and Meng et al. (2014) similarity measures; (4) the Wu and Palmer (1994) similarity measure obtains a statistically significant lower average Spearman correlation value than all of the remaining measures; (5) our *coswJ&C* similarity measure and the measure introduced by Zhou et al. (2008b) obtain a statistically significant higher average Spearman correlation value than all of the classic IC-based measures, whilst our *cosJ&C* measure and the Garla and Brandt (2012) measure statistically outperform the Resnik and Lin similarity measures; and finally, (6) the Rada et al. (1989) similarity measure and all measures derived from it, such as the measures introduced by Leacock and Chodorow (1998) and Pedersen et al. (2007), together with the Al-Mubaid and Nguyen (2009) measure, are only outperformed in a statistically significant manner by our *coswJ&C* similarity measure, and the measures introduced by Zhou et al. (2008b) and Meng et al. (2014).

In summary, conclusions (1) and (2) above prove hypothesis H1 on the outperformance of the path-based similarity measures by a group of state-of-the-art IC-based similarity measures. Conclusion (5) above proves the hypothesis H4 on the outperformance of the classic IC-based similarity measures by a small set of state-of-the-art methods. On the other hand, the conclusion (6) above is very significant because it proves for the first time that *only this small set of state-of-the-art IC-based similarity measures have been able to obtain a statistically significant higher average Spearman correlation value than the family of path-based similarity measures.* If we reproduce the statistical significance analysis in table 14 using the average Pearson correlation as sample set, we could see that most IC-based similarity measures obtain a statistically significant higher average Pearson correlation than the path-based measures, a fact that endorses the common belief that the path-based similarity measures have been definitively outperformed by the

family of IC-based similarity measures. However, the results shown in table 14 reopen the debate. We argue that the lack of a statistically significant difference between the Garla and Brandt (2012) and Pedersen et al. (2007) similarity measures, and thus any other measure derived from Rada et al. (1989), is mainly responsible for the lack of a statistically significant difference in performance reported by Alonso and Contreras (2016) for the use of the two aforementioned measures in a biomedical IR task. The latter facts endorse our idea that research into the area should focus on the improvement in the performance based on the Spearman rank correlation, because this latter metric could predict the expected performance in applications based on similarity measures better.

We note other significant fact. Our *coswJ&C* similarity measure, and the measures introduced by Zhou et al. (2008b) and Meng et al. (2014), are all hybrid IC-based similarity measures that integrate an IC model with any path-based feature. Among the latter aforementioned measures, the *coswJ&C* similarity measure is the only one that defines a real IC-based weighted graph, whilst the other two measures integrate a pure edge-counting measure in their formulas. Our experimental results and the significance analysis show that the IC-based weighted distance on a taxonomy, as proposed by the *coswJ&C* similarity measure, is currently the best approach for maximizing the Spearman rank correlation value, thus, this type of taxonomical feature should be explored in future developments into ontology-based similarity measures, despite its high computational cost.

6.5 Impact of the IC models on the similarity measures

The last four rows in tables 10 and 11 show a set of statistics considering the Pearson and Spearman correlation values reported by each similarity measure (column) as a random variable evaluated on all IC models. These statistics allow the following conclusions to be drawn: (1) most IC-based similarity measures exhibit a moderate standard deviation in the Pearson and Spearman correlation values as regard the set of IC models; (2) most IC-based similarity measures in table 11 exhibit a peak ratio greater than 1.0 times their standard deviation, a fact that supports our H6 hypothesis which states that most IC-based similarity measures perform better with a specific IC model; and (3) the standard deviation of the Spearman correlation of the IC-based similarity measures as regards the IC models is statistically significant lower than the standard deviation of the Pearson correlation, a fact that is supported by a p-value of 0.0073 between both random sets. This latter fact means that the performance of the IC-based similarity measures as a function of the IC models is more stable in terms of the Spearman rank correlation than the Pearson metric.

We conclude that every similarity measure should be used with its best performing IC model in any practical application. However, there is no strong evidence confirming that the outperformance of a similarity measure in any word similarity benchmark can be extrapolated to

other applications (see our discussion in section 3). Our most significant conclusion as regards the IC models is as follows: the two best performing and preferred IC models by most IC-based similarity measures, and thus, the most practical IC models, are those introduced by Sánchez et al. (2011) and Seco et al. (2004).

6.6 New state-of-the-art results

The new state-of-the-art in intrinsic IC models and intrinsic IC-based similarity measures is set by the Sánchez et al. (2011) IC model in combination with our *coswJ&C* similarity measure, and the Seco et al. (2004) IC model in combination with the Zhou et al. (2008b) similarity measure. Likewise, these two latter intrinsic IC-based similarity measures obtain a statistically significant higher performance than the remaining methods. Thus, the four aforementioned methods are convincing winners among the families of IC models and ontology-based similarity measures. The *coswJ&C* similarity measure obtains the highest average Spearman correlation value and the highest overall averaged Pearson-Spearman correlation value on all datasets, as well as the highest Spearman correlation value in four of the five datasets evaluated, and the highest Pearson correlation values in the RG65 and MC28 datasets. On the other hand, the Zhou et al. (2008b) similarity measure obtains the highest average Pearson correlation value on all datasets and the highest Spearman correlation value in the SimLex665 dataset.

The set of classic IC-based similarity measures, defined by the Resnik, Lin and Jiang-Conrath measures, have also been definitively outperformed in a statistically significant manner by a small set of IC-based similarity measures developed during the last decade, among which we find the similarity measures introduced by Zhou et al. (2008b) and the *coswJ&C* measure introduced by Lastra-Díaz and García-Serrano (2015b). In addition, the J&C similarity measure and its two monotone transformations, our *cosJ&C* measure and the Garla and Brandt (2012) similarity measure, obtain a statistically significant higher average Spearman correlation than the Resnik and Lin similarity measures, and the *cosJ&C* obtains a statistically significant average Pearson correlation value than the J&C similarity measure. However, we also prove that there is no a statistically significant difference between the two aforementioned pairs of outperforming IC-based similarity measures.

According to the results obtained, the two similarity measures with the best overall performance are the two hybrid IC-based similarity measures defined by the *coswJ&C* introduced by Lastra-Díaz and García-Serrano (2015b) and the Zhou et al. (2008b) measure. However, their computational cost prevents their practical use in comparison with other measures, such as the *cosJ&C* introduced by Lastra-Díaz and García-Serrano (2015b) and the Hadj Taieb et al. (2014b) measure. There is no statistically significant difference between these two latter measures. The *cosJ&C* measure obtains a higher Spearman correlation on average than the Hadj Taieb et al. (2014b) measure, whilst the Hadj Taieb et al. (2014b) measure obtains a higher Pearson cor-

relation on average than the previous one. Thus, the *cosJ&C* and [Hadj Taieb et al. \(2014b\)](#) measures are, statistically speaking, the best option from the aforementioned set of similarity measures with a practical computational cost.

6.7 Monotone transformations.

The Spearman rank correlation value is invariant to monotone transformations from any similarity measure, thus, its exhaustive evaluation for all the similarity measures and IC models has confirmed that a lot of similarity measures are monotone transformations of other classic measures, as well as the findings of other unknown cases. For instance, the Spearman correlation metric reported by the *FaITH* similarity measure introduced by [Pirró and Euzenat \(2010\)](#) reveals that it is a monotone transformation of the Lin measure like the measure introduced by [Meng and Gu \(2012\)](#). Indeed, there are many cases like these. For instance, the similarity measure introduced by [Leacock and Chodorow \(1998\)](#), the *sim_{Path}* measure of [Pedersen et al. \(2007\)](#), and the *sim_{Li_s3}* measure of [Li et al. \(2003\)](#), all which are monotone transformations of the [Rada et al. \(1989\)](#) measure, whilst the *sim_{path_IC}* measure of [Garla and Brandt \(2012\)](#) and the *sim_{cosJ&C}* measure introduced by [Lastra-Díaz and García-Serrano \(2015b\)](#) are monotone transformations of the J&C similarity measure as defined in table 2. We confirmed experimentally that in all of the aforementioned cases, the transformed measures preserve the Spearman correlation values obtained by their respective base measures, differing only in their Pearson correlation values. Table 3 shows a factorization of the latter similarity measures that proves the aforementioned monotonicity relationships.

As a consequence of the aforementioned monotonicity relationships, there is a reduced number of different strategies to estimate the degree of similarity using an ontology-based similarity measure, despite many similarity measures having been proposed in the literature. We argue that the monotonicity relationships between a large set of similarity measures are the main cause behind the lack of a statistically significant difference between most of the similarity measures evaluated herein. Thus, the research community should explore either new measurement methods or new similarity models in order to bring about significant progress in the state of the problem. On the other hand, the results obtained by the measures introduced by [Meng and Gu \(2012\)](#), [Garla and Brandt \(2012\)](#), [Pirró and Euzenat \(2010\)](#) and [Lastra-Díaz and García-Serrano \(2015b\)](#), prove that a proper scaling and normalization of the similarity measures is a good strategy to improve the Pearson correlation metric slightly. Therefore, the research should focus on the search for a significant improvement in the Spearman correlation metric, which is also closely related to the measurement strategy and similarity model used.

6.8 Computational complexity

Table 9 compares the running time of each similarity measure in the evaluation of the MC28 dataset. The

feature-based measure of [Sánchez et al. \(2012\)](#) obtains the lowest running time, making it the fastest among all of the measures. As we expected from an analysis of their definitions, all non hybrid IC-based similarity measures obtain a running time that is almost identical to that reported by the Resnik measure defined as baseline. The small differences are only attributable to the activity of the operating system during the experiments, because these IC-based similarity measures share the same IC-based factors. On the other hand, the hybrid IC-based similarity measures exhibit a running time of between 52 and 237 times greater than the baseline, making our *coswJ&C* similarity measure the slowest among all of the measures. Thus, the computational complexity of the hybrid IC-based measures is roughly two orders of magnitude greater than the complexity of the remaining IC-based similarity measures. Despite all hybrid IC-based similarity measures using the same implementation of the Dijkstra algorithm in our software library, our *coswJ&C* similarity measure requires the measurement of the length of the shortest path between concepts on a non-uniform and real-valued weighted graph, whilst the rest of the hybrid IC-based similarity measures only require the edge count to be obtained, thus, the Dijkstra algorithm is much faster in this latter case.

6.9 Confirming our hypotheses

The hypotheses H1, H2, H3, H4, H5, H6 and H8 introduced in section 1.2 have been positively confirmed by the data obtained from our experiments, they having been answered in the discussion above. Finally, hypothesis H7 on the outperformance of the state-of-the-art IC-based similarity measures on the best corpus-based similarity measures in the SimLex666 dataset, is also confirmed by comparing the best Pearson and Spearman correlation values obtained by most IC-based similarity measures in tables 23 and 24, with the results for these metrics reported for the best corpus-based method in the SimLex dataset (Pearson=0.599, Spearman=0.591), as reported in a recent benchmark by [Banjade et al. \(2015\)](#).

6.10 Contradictory results

We obtained several contradictory results in our experiments, confirming the same findings reported in our aforementioned works, as well as other new ones. For instance, [Meng and Gu \(2012\)](#) and [Meng et al. \(2014\)](#) report Pearson correlation values of 0.8804 and 0.8817 respectively with the [Seco et al. \(2004\)](#) IC model in the RG65 dataset, whilst we obtained 0.8596 and 0.8486 respectively. [Gao et al. \(2015\)](#) report a Pearson correlation value of 0.885 for their similarity measure in the RG65 dataset with an unknown corpus-based IC model, whilst we obtained 0.87098 herein. [Adhikari et al. \(2015\)](#) report the following Pearson correlation values of 0.86, 0.86 and 0.84 for their IC model in the MC30 dataset with the Resnik, Lin and Jiang-Conrath similarity measure respectively, whilst we obtained 0.8211, 0.8410 and 0.8331 in the MC28 dataset. These facts confirm the reproducibility problems in the area. Thus, we invite the

research community to reproduce the methods and experiments reported in the literature in order to confirm or refute the results reported herein.

7 Conclusions and future work

We have introduced a refinement of our recent family of well-founded Information Content models, eight new intrinsic IC models and one new corpus-based IC model and a very detailed experimental survey on WordNet. We have proven that the proposed refinement improves the performance of our family of well-founded IC models, and six of our new IC models obtain rivaling results as regard the state-of-the-art intrinsic IC models, making the new *CondProbRefHyponyms* and *CondProbRefCosine* IC models our best performing IC models.

The [Seco et al. \(2004\)](#) and [Sánchez et al. \(2011\)](#) IC models set the state of the art for the IC models, and the [Seco et al. \(2004\)](#), [Sánchez et al. \(2011\)](#) and [Yuan et al. \(2013\)](#) IC models are the only intrinsic IC models that statistically outperform the best performing corpus-based IC model used as baseline. However, we prove that there is no statistically significant difference between most intrinsic IC models and the corpus-based Resnik IC model defined as baseline. Therefore, the aforementioned set of intrinsic IC models can be considered as a practical alternative to the corpus-based ones, and they should be selected in accordance with the IC-based similarity measure used. On the other hand, the detailed experiment survey carried-out herein allows a very significant conclusion to be drawn: despite the research effort made during the last decade, the [Seco et al. \(2004\)](#) IC model is still the state of the art on average.

The new state-of-the-art in intrinsic IC models and intrinsic IC-based similarity measures is set by the [Sánchez et al. \(2011\)](#) IC model in combination with our *coswJ&C* similarity measure, and the [Seco et al. \(2004\)](#) IC model in combination with the [Zhou et al. \(2008b\)](#) similarity measure. The set of classic IC-based similarity measures, defined by the Resnik, Lin and Jiang-Conrath measures, have also been definitively outperformed in a statistically significant manner by a small set of IC-based similarity measures developed during the last decade, among which we find the similarity measures introduced by [Zhou et al. \(2008b\)](#) and the *coswJ&C* introduced by [Lastra-Díaz and García-Serrano \(2015b\)](#). In addition, the J&C similarity measure and its two monotone transformations, our *cosJ&C* measure and the [Garla and Brandt \(2012\)](#) similarity measure, statistically outperform the Resnik and Lin similarity measures, and the *cosJ&C* similarity measure obtains a statistically significant higher average Pearson correlation value than the J&C similarity measure. However, we also prove that there is no a statistically significant difference between the two aforementioned pairs of outperforming IC-based similarity measures.

Despite our *coswJ&C* similarity measure and the [Zhou et al. \(2008b\)](#) measure setting the state of the art of the problem, their computational cost prevent their practical use in comparison with other measures, such as the *cosJ&C* introduced by [Lastra-Díaz and García-](#)

[Serrano \(2015b\)](#) and the [Hadj Taieb et al. \(2014b\)](#) measure. There is no a statistically significant difference between the two latter aforementioned measures. Thus, the *cosJ&C* and [Hadj Taieb et al. \(2014b\)](#) measures are, statistically speaking, the best option from the aforementioned set of similarity measures with a practical computational cost.

We have proven that the state of the art in ontology-based similarity measures and concept similarity models is led by the family of IC-based measures, more specifically by the measures derived from the Jiang-Conrath similarity measure. In addition, we have made another significant finding. Contrary to the common belief among the research community, only a small set of state-of-the-art hybrid IC-based similarity measures derived from the J&C measure obtain a statistically significant higher average Spearman correlation value than the family of path-based similarity measures, a fact that explains some unexpected results in applications based on similarity measures reported in the literature, such as that reported by [Alonso and Contreras \(2016\)](#).

Finally, as forthcoming activities, we are going to introduce and releasing HESML in a forthcoming paper [Lastra-Díaz and García-Serrano \(2016\)](#), which is a new scalable Java software library of ontology-based semantic similarity measures and IC models. In addition, HESML will be released with a replication dataset called *WN-SimRep v1*, as well as a set of reproducible experiments which allow automatically reproducing all the results reported in our two aforementioned works and herein. The aforementioned forthcoming paper is part of a novel initiative on computational reproducibility recently introduced by [Chirigati et al. \(2016\)](#), whose pioneering work is introduced by [Wolke et al. \(2016\)](#) with the aim of aiding the exact replication of several dynamic resource allocation strategies in cloud data centers evaluated in another companion paper [Wolke et al. \(2015\)](#). Our reproducible experiments are based on ReprZip, which is a virtualization tool introduced by [Chirigati et al. \(2013b\)](#) and [Chirigati et al. \(2013a\)](#), whose aim is to warrant the exact replication of experimental results onto a different system from that originally used into their creation.

8 Acknowledgements

Ted Pedersen kindly answered all our questions and provided us the WordNet-based frequency files used to build the corpus-based IC models included in our experiments. Mohamed Hadj Taieb kindly offered us his total support in replicating their similarity measures exactly. Emmanuel Pothos, Vijay Garla, Abdulgabbar Saif, Jorge Martínez-Gil and Lubomir Stanchev kindly answered our questions about their works. Mark Hallett checked the proper use of English. To all of them, we would like to express our most sincere gratitude. Finally, we also express our gratitude to the anonymous reviewers by their remarks in order to improve the quality of this work. This work has been partially supported by the Spanish VOXPOPULI (TIN2013-47090-C3-1-P) Project.

9 Appendix

See summary tables 10, 11, 12, 13 and 14. All raw data resulting from the evaluation is shown in tables 15 to 26 next the bibliography.

References

- Adhikari, A., Singh, S., Dutta, A., Dutta, B., Nov. 2015. A novel information theoretic approach for finding semantic similarity in WordNet. In: Proceedings of IEEE International Technical Conference (TENCON-2015). IEEE, Macau, China, pp. 1–6.
- Agirre, E., Alfonseca, E., Hall, K., Kravalova, J., Pasca, M., Soroa, A., 2009. A Study on Similarity and Relatedness Using Distributional and WordNet-based Approaches. In: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics. NAACL '09. ACL, Stroudsburg, PA, USA, pp. 19–27.
- Al-Mubaid, H., Nguyen, H. A., Jul. 2009. Measuring Semantic Similarity Between Biomedical Concepts Within Multiple Ontologies. IEEE transactions on systems, man and cybernetics. Part C, Applications and reviews: a publication of the IEEE Systems, Man, and Cybernetics Society 39 (4), 389–398.
- Alonso, I., Contreras, D., Feb. 2016. Evaluation of semantic similarity metrics applied to the automatic retrieval of medical documents: an UMLS approach. Expert Systems with Applications 44, 386–399.
- Alvarez, M. A., Lim, S., Sep. 2007. A Graph Modeling of Semantic Similarity between Words. In: Proceedings of the First IEEE International Conference on Semantic Computing (ICSC 2007). IEEE Computer Society, Irvine, California, USA, pp. 355–362.
- Aouicha, M. B., Taieb, M. A. H., 17 Dec. 2015. Computing semantic similarity between biomedical concepts using new information content approach. Journal of Biomedical Informatics.
- Ash, R. B., Doléans-Dade, C. A., 2000. Probability & Measure Theory, 2nd Edition. Academic Press.
- Banjade, R., Maharjan, N., Niraula, N. B., Rus, V., Gautam, D., 14 Apr. 2015. Lemon and Tea Are Not Similar: Measuring Word-to-Word Similarity by Combining Different Methods. In: Gelbukh, A. (Ed.), Proceedings of the 16th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing). Vol. 9041 of LNCS. Springer, Cayro, Egypt, pp. 335–346.
- Batet, M., Harispe, S., Ranwez, S., Sánchez, D., Ranwez, V., 1 Nov. 2014. An information theoretic approach to improve semantic similarity assessments across multiple ontologies. Information sciences 283, 197–210.
- Batet, M., Sánchez, D., Valls, A., Feb. 2011. An ontology-based measure to compute semantic similarity in biomedicine. Journal of Biomedical Informatics 44 (1), 118–125.
- Ben Aouicha, M., Hadj Taieb, M. A., Ezzeddine, M., Apr. 2016a. Derivation of “is a” taxonomy from Wikipedia Category Graph. Engineering Applications of Artificial intelligence 50, 265–286.
- Ben Aouicha, M., Taieb, M. A. H., Ben Hamadou, A., 28 Mar. 2016b. Taxonomy-based information content and wordnet-wiktionary-wikipedia glosses for semantic relatedness. Applied Intelligence, 1–37.
- Blanchard, E., Harzallah, M., Kuntz, P., 2008. A generic framework for comparing semantic similarities on a subsumption hierarchy. In: Ghallab, M., Spyropoulos, C. D., Fakotakis, N., Avouris, N. (Eds.), Proceedings of the ECAI. Vol. 178 of Frontiers in Artificial Intelligence and Applications. IOS Press, pp. 20–24.
- Budanitsky, A., Hirst, G., 2001. Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. In: Workshop on WordNet and Other Lexical Resources, Second Meeting of the North American Chapter of the Association for Computational Linguistics. Vol. 2. ACL, Pittsburgh, PA, pp. 29–34.
- Budanitsky, A., Hirst, G., Mar. 2006. Evaluating WordNet-based Measures of Lexical Semantic Relatedness. Computational Linguistics 32 (1), 13–47.
- Busemeyer, J. R., Bruza, P. D., 2012. Quantum models of cognition and decision. Cambridge University Press.
- Chaves-González, J. M., Martínez-Gil, J., Jan. 2013. Evolutionary algorithm based on different semantic similarity functions for synonym recognition in the biomedical domain. Knowledge-Based Systems 37, 62–69.
- Chiang, J.-H., Ho, S.-H., Wang, W.-H., Oct. 2008. Similar genes discovery system (SGDS): Application for predicting possible pathways by using GO semantic similarity measure. Expert Systems with Applications 35 (3), 1115–1121.
- Chirigati, F., Capone, R., Rampin, R., Freire, J., Shasha, D., Mar. 2016. A collaborative approach to computational reproducibility. Information Systems 59, 95–97.
- Chirigati, F., Shasha, D., Freire, J., 2013a. Packing Experiments for Sharing and Publication. In: Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data. SIGMOD '13. ACM, New York, NY, USA, pp. 977–980.
- Chirigati, F., Shasha, D., Freire, J., 2013b. Rezip: Using provenance to support computational reproducibility. In: Presented as part of the 5th USENIX Workshop on the Theory and Practice of Provenance. usenix.org.

- Couto, F. M., Pinto, H. S., Oct. 2013. The next generation of similarity measures that fully explore the semantics in biomedical ontologies. *Journal of Bioinformatics and Computational Biology* 11 (5), 1371001.
- Couto, F. M., Silva, M. J., Coutinho, P. M., Apr. 2007. Measuring semantic similarity between Gene Ontology terms. *Data & Knowledge Engineering* 61 (1), 137–152.
- Cross, V., Hu, X., 2011. Using Semantic Similarity in Ontology Alignment. In: *Proceedings of the Sixth International Workshop on Ontology Matching (OM), 10th Int. Semantic Web Conference (ISWC 2011)*. Bonn Germany, pp. 61–72.
- Dagher, G. G., Fung, B. C. M., Jul. 2013. Subject-based semantic document clustering for digital forensic investigations. *Data & Knowledge Engineering* 86, 224–241.
- Fernando, S., Stevenson, M., 2008. A semantic similarity approach to paraphrase detection. In: *Proceedings of the 11th Annual Research Colloquium of the UK Special Interest Group for Computational Linguistics*. Oxford, UK, pp. 45–52.
- Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., Ruppin, E., 1 Jan. 2002. Placing search in context: the concept revisited. *ACM Transactions on Information Systems (TOIS)* 20 (1), 116–131.
- Fokkens, A., Van Erp, M., Postma, M., Pedersen, T., Vossen, P., Freire, N., 4 Aug. 2013. Offspring from Reproduction Problems: What Replication Failure Teaches Us. In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*. ACL, Sofia, Bulgaria, pp. 1691–1701.
- Gabrilovich, E., Markovitch, S., 2007. Computing Semantic Relatedness Using Wikipedia-based Explicit Semantic Analysis. In: *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI-07)*. Vol. 7. Morgan Kaufmann Publishers Inc., Hyderabad, India, pp. 1606–1611.
- Gao, J. B., Zhang, B. W., Chen, X. H., Mar. 2015. A WordNet-based semantic similarity measurement combining edge-counting and information content theory. *Engineering Applications of Artificial Intelligence* 39, 80–88.
- Garla, V. N., Brandt, C., 10 Oct. 2012. Semantic similarity in the biomedical domain: an evaluation across knowledge sources. *BMC bioinformatics* 13:261.
- Hadj Taieb, M. A., Ben Aouicha, M., Ben Hamadou, A., 1 Nov. 2014a. A new semantic relatedness measurement using WordNet features. *Knowledge and Information Systems* 41 (2), 467–497.
- Hadj Taieb, M. A., Ben Aouicha, M., Ben Hamadou, A., Nov. 2014b. Ontology-based approach for measuring semantic similarity. *Engineering Applications of Artificial Intelligence* 36, 238–261.
- Hadj Taieb, M. A., Ben Aouicha, M., Bourouis, Y., 22 Jun. 2015. Fm3s: Features-based measure of sentences semantic similarity. In: Onieva, E., Santos, I., Osaba, E., Quintián, H., Corchado, E. (Eds.), *Proceedings of the 10th International Conference on Hybrid Artificial Intelligent Systems (HAIS 2015)*. Vol. 9121 of LNCS. Springer, Bilbao, Spain, pp. 515–529.
- Harispe, S., Imoussaten, A., Troussel, F., Montmain, J., Aug. 2015a. On the consideration of a bring-to-mind model for computing the Information Content of concepts defined into ontologies. In: *Proceedings of the IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2015)*. IEEE, Istanbul, Turkey, pp. 1–8.
- Harispe, S., Ranwez, S., Janaqi, S., Montmain, J., May 2015b. Semantic Similarity from Natural Language and Ontology Analysis. Vol. 8 of *Synthesis Lectures on Human Language Technologies*. Morgan & Claypool publishing.
- Hill, F., Reichart, R., Korhonen, A., 2015. SimLex-999: Evaluating Semantic Models with (Genuine) Similarity Estimation. *Computational Linguistics* 41 (4), 665–695.
- Hirst, G., St-Onge, D., 1998. Lexical chains as representations of context for the detection and correction of malapropisms. In: Fellbaum, C. (Ed.), *WordNet: An electronic lexical database*. Massachusetts Institute of Technology, pp. 305–332.
- Hughes, T., Ramage, D., 28 Jun. 2007. Lexical Semantic Relatedness with Random Graph Walks. In: *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*. ACL, Prague, Czech Republic, pp. 581–589.
- Jeong, B., Lee, D., Cho, H., Lee, J., Apr. 2008. A novel method for measuring semantic similarity for xml schema matching. *Expert Systems with Applications* 34 (3), 1651–1658.
- Jiang, J. J., Conrath, D. W., 1997. Semantic similarity based on corpus statistics and lexical taxonomy. In: *Proceedings of International Conference Research on Computational Linguistics (ROCLING X)*. pp. 19–33.
- Jurgens, D., Pilehvar, M. T., Navigli, R., Oct. 2015. Cross level semantic similarity: an evaluation framework for universal measures of similarity. *Language Resources and Evaluation*, 1–29.
- Lastra-Díaz, J. J., 29 Sep. 2014. Intrinsic Semantic Spaces for the representation of documents and semantic annotated data. Department of Computer Languages and Systems. Universidad Nacional de Educación a Distancia (UNED). <http://e-spacio.uned.es/fez/view/bibliuned:master-ETSIInformatica-LSI-Jlastra>.
- Lastra Díaz, J. J., García Serrano, A., 19 Dec. 2014. System and method for the indexing and retrieval

- of semantically annotated data using an ontology-based information retrieval model. United States Patent and Trademark Office (USPTO) application US14/576,679.
- Lastra-Díaz, J. J., García-Serrano, A., Nov. 2015a. A new family of information content models with an experimental survey on WordNet. *Knowledge-Based Systems* 89, 509–526.
- Lastra-Díaz, J. J., García-Serrano, A., Nov. 2015b. A novel family of IC-based similarity measures with a detailed experimental survey on WordNet. *Engineering Applications of Artificial Intelligence Journal* 46, 140–153.
- Lastra-Díaz, J. J., García-Serrano, A., 2016. HESML: a scalable ontology-based semantic similarity measures library with a set of reproducible experiments and a replication dataset. To appear in *Information Systems journal*.
- Le, M., Fokkens, A., 7 Sep. 2015. Taxonomy Beats Corpus in Similarity Identification, but Does It Matter? In: Angelova, G., Bontcheva, K., Mitkov, R. (Eds.), *Proceedings of International Conference on Recent Advances in Natural Language Processing*. Hissar, Bulgaria, pp. 346–354.
- Leacock, C., Chodorow, M., 1998. Combining local context and WordNet similarity for word sense identification. In: Fellbaum, C. (Ed.), *WordNet: An electronic lexical database*. MIT Press, Ch. 11, pp. 265–283.
- Lee, M. C., May 2011. A novel sentence similarity measure for semantic-based expert systems. *Expert Systems with Applications* 38 (5), 6392–6399.
- Levy, R., Ein-Dor, L., Hummel, S., Rinott, R., Slonim, N., 26 Jul. 2015. TR9856: A Multi-word Term Relatedness Benchmark. In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Short Papers)*. ACL, Beijing, China, pp. 419–424.
- Li, P., Wang, H., Zhu, K. Q., Wang, Z., Hu, X.-G., Wu, X., 2015. A Large Probabilistic Semantic Network based Approach to Compute Term Similarity. *IEEE Transactions on Knowledge and Data Engineering* 27 (10), 2604–2617.
- Li, Y., Bandar, Z. A., McLean, D., 2003. An approach for measuring semantic similarity between words using multiple information sources. *IEEE Transactions on Knowledge and Data Engineering* 15 (4), 871–882.
- Lin, D., 1998. An information-theoretic definition of similarity. In: *Proceedings of the 15th International Conference on Machine Learning*. Vol. 98. Madison, WI, pp. 296–304.
- Liu, M., Shen, W., Hao, Q., Yan, J., Dec. 2009. An weighted ontology-based semantic similarity algorithm for web service. *Expert Systems with Applications* 36 (10), 12480–12490.
- Martínez, S., Sánchez, D., Valls, A., 27 Oct. 2010. Ontology-based anonymization of categorical values. In: *Modeling Decisions for Artificial Intelligence*. Vol. 6408 of LNCS. Springer Berlin Heidelberg, pp. 243–254.
- Martinez-Gil, J., Dec. 2016. CoTO: A Novel Approach for Fuzzy Aggregation of Semantic Similarity Measures. *Cognitive Systems Research* 40, 8–17.
- McInnes, B. T., Pedersen, T., Dec. 2013. Evaluating measures of semantic similarity and relatedness to disambiguate terms in biomedical text. *Journal of biomedical informatics* 46 (6), 1116–1124.
- Meijer, K., Frasincar, F., Hogenboom, F., Jun. 2014. A semantic approach for extracting domain taxonomies from text. *Decision Support Systems* 62, 78–93.
- Meng, L., Gu, J., 2012. A New Model for Measuring Word Sense Similarity in WordNet. In: *Proceedings of the 4th International Conference on Advanced Communication and Networking, ASTL*. Vol. 14. pp. 18–23.
- Meng, L., Gu, J., Zhou, Z., Sep. 2012. A new model of information content based on concept’s topology for measuring semantic similarity in WordNet. *International Journal of Grid and Distributed Computing* 5 (3), 81–93.
- Meng, L., Huang, R., Gu, J., Jun. 2014. Measuring Semantic Similarity of Word Pairs Using Path and Information Content. *International Journal of Future Generation Communication & Networking* 7 (3), 183–194.
- Mihalcea, R., Corley, C., Strapparava, C., 2006. Corpus-based and knowledge-based measures of text semantic similarity. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 1. AAAI Press, pp. 775–780.
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J., 2013. Distributed representations of words and phrases and their compositionality. In: Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z., Weinberger, K. Q. (Eds.), *Advances in Neural Information Processing Systems 26 (NIPS 2013)*. NIPS Foundation, Inc., pp. 3111–3119.
- Miller, G. A., 1995. WordNet: A Lexical Database for English. *Communications of the ACM* 38 (11), 39–41.
- Miller, G. A., Charles, W. G., 1991. Contextual correlates of semantic similarity. *Language and cognitive processes* 6 (1), 1–28.
- Mohammad, S., Hirst, G., 2006. Distributional Measures of Concept-distance: A Task-oriented Evaluation. In: *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing. EMNLP ’06*. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 35–43.
- Mohammad, S. M., Hirst, G., 8 Mar. 2012. Distributional Measures of Semantic Distance: A Survey. arXiv:1203.1858.

- Montani, S., Leonardi, G., Quaglini, S., Cavallini, A., Micieli, G., 1 Jun. 2015. A knowledge-intensive approach to process similarity calculation. *Expert Systems with Applications* 42 (9), 4207–4215.
- Oliva, J., Serrano, J. I., del Castillo, M. D., Iglesias, A., Apr. 2011. SyMSS: A syntax-based measure for short-text semantic similarity. *Data & Knowledge Engineering* 70 (4), 390–405.
- Patwardhan, S., Banerjee, S., Pedersen, T., Feb. 2003. Using Measures of Semantic Relatedness for Word Sense Disambiguation. In: Gelbukh, A. (Ed.), *Proceedings of the 4th International Conference on Computational Linguistics and Intelligent Text Processing (CICLING 2003)*. Vol. 2588 of LNCS. Springer, Mexico D.F., pp. 241–257.
- Patwardhan, S., Pedersen, T., 2006. Using WordNet-based context vectors to estimate the semantic relatedness of concepts. In: *Proceedings of the EACL 2006 Workshop Making Sense of Sense-Bringing Computational Linguistics and Psycholinguistics Together*. Vol. 1501. Trento, Italy, pp. 1–8.
- Pedersen, T., 2008a. Empiricism Is Not a Matter of Faith. *Computational Linguistics* 34 (3), 465–470.
- Pedersen, T., 2008b. WordNet-InfoContent-3.0.tar dataset repository. https://www.researchgate.net/publication/273885902_WordNet-infoContent-3.0.tar.
- Pedersen, T., 2010. Information Content Measures of Semantic Similarity Perform Better Without Sense-tagged Text. In: *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. HLT '10*. ACL, Stroudsburg, PA, USA, pp. 329–332.
- Pedersen, T., 25 Nov. 2013. Measuring the Similarity and Relatedness of Concepts: a MICAI 2013 Tutorial. Invited talk in the 12th Mexican International Conference on Artificial Intelligence (MICAI 2013). <http://dx.doi.org/10.13140/RG.2.1.3025.6164>.
- Pedersen, T., Pakhomov, S. V. S., Patwardhan, S., Chute, C. G., Jun. 2007. Measures of semantic similarity and relatedness in the biomedical domain. *Journal of biomedical informatics* 40 (3), 288–299.
- Pennington, J., Socher, R., Manning, C. D., 2014. Glove: Global vectors for word representation. *Proceedings of the Empirical Methods in Natural Language Processing (EMNLP 2014)* 12, 1532–1543.
- Pesquita, C., Faria, D., Falcao, A. O., Lord, P., Couto, F. M., 2009. Semantic similarity in biomedical ontologies. *PLoS computational biology* 5 (7), e1000443.
- Pilehvar, M. T., Navigli, R., Nov. 2015. From senses to texts: An all-in-one graph-based approach for measuring semantic similarity. *Artificial Intelligence* 228, 95–128.
- Pirró, G., Nov. 2009. A semantic similarity metric combining features and intrinsic information content. *Data & Knowledge Engineering* 68 (11), 1289–1308.
- Pirró, G., Euzenat, J., 7 Nov. 2010. A Feature and Information Theoretic Framework for Semantic Similarity and Relatedness. In: Patel-Schneider, P. F., Pan, Y., Hitzler, P., Mika, P., Zhang, L., Pan, J. Z., Horrocks, I., Glimm, B. (Eds.), *Proceedings of the 9th International Semantic Web Conference, ISWC 2010*. Vol. 6496 of LNCS. Springer, Shanghai, China, pp. 615–630.
- Pirró, G., Seco, N., Jan. 2008. Design, Implementation and Evaluation of a New Semantic Similarity Metric Combining Features and Intrinsic Information Content. In: Meersman, R., Tari, Z. (Eds.), *On the Move to Meaningful Internet Systems: OTM 2008*. Vol. 5332 of LNCS. Springer, pp. 1271–1288.
- Pirró, G., Talia, D., May 2010. Ufome: An ontology mapping system with strategy prediction capabilities. *Data & Knowledge Engineering* 69 (5), 444–471.
- Pothos, E. M., Barque-Duran, A., Yearsley, J. M., Trueblood, J. S., Busemeyer, J. R., Hampton, J. A., 25 Feb. 2015. Progress and current challenges with the quantum similarity model. *Frontiers in psychology* 6, 205.
- Pothos, E. M., Busemeyer, J. R., Trueblood, J. S., Jul. 2013. A quantum geometric model of similarity. *Psychological review* 120 (3), 679–696.
- Pothos, E. M., Trueblood, J. S., Feb. 2015. Structured representations in a quantum probability model of similarity. *Journal of Mathematical Psychology* 64–65, 35–43.
- Rada, R., Mili, H., Bicknell, E., Blettner, M., 1989. Development and application of a metric on semantic nets. *IEEE Transactions on Systems, Man, and Cybernetics* 19 (1), 17–30.
- Ramage, D., Rafferty, A. N., Manning, C. D., 2009. Random Walks for Text Semantic Similarity. In: *Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing. TextGraphs-4*. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 23–31.
- Resnik, P., 20 Aug. 1995. Using information content to evaluate semantic similarity in a taxonomy. In: *Proceedings of the International Joint Conferences on Artificial Intelligence (IJCAI 1995)*. Vol. 1. Montreal, Canada, pp. 448–453.
- Resnik, P., Jul. 1999. Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. *Journal of Artificial Intelligence Research* 11, 95–130.
- Rubenstein, H., Goodenough, J. B., Oct. 1965. Contextual Correlates of Synonymy. *Communications of the ACM* 8 (10), 627–633.

- Saif, A., Ab Aziz, M. J., Omar, N., 2014. Evaluating Knowledge-Based Semantic Measures on Arabic. *International Journal on Communications Antenna and Propagation (IRECAP)* 4 (5), 180–194.
- Sánchez, D., Batet, M., Oct. 2011. Semantic similarity estimation in the biomedical domain: an ontology-based information-theoretic perspective. *Journal of biomedical informatics* 44 (5), 749–759.
- Sánchez, D., Batet, M., 2012. A new model to compute the information content of concepts from taxonomic knowledge. *International Journal on Semantic Web and Information Systems (IJSWIS)* 8 (2), 34–50.
- Sánchez, D., Batet, M., Isern, D., Mar. 2011. Ontology-based information content computation. *Knowledge-Based Systems* 24 (2), 297–303.
- Sánchez, D., Batet, M., Isern, D., Valls, A., Jul. 2012. Ontology-based semantic similarity: A new feature-based approach. *Expert Systems with Applications* 39 (9), 7718–7728.
- Sanfilippo, A., Tratz, S., Gregory, M., Chappell, A., Whitney, P., Posse, C., Paulson, P., Baddeley, B., Hohimer, R., White, A., 7 Nov. 2005. Ontological annotation with wordnet. In: *Proceedings of the 5th International Workshop on Knowledge Markup and Semantic Annotation (SemAnnot 2005)* located at the 4rd International Semantic Web Conference (ISWC 2005). Galway, Ireland, pp. 27–36.
- Sebti, A., Barfroush, A. A., Oct. 2008. A new word sense similarity measure in WordNet. In: *Proceedings of the International Multiconference on Computer Science and Information Technology, IMCSIT 2008*. IEEE, pp. 369–373.
- Seco, N., Veale, T., Hayes, J., 2004. An intrinsic information content metric for semantic similarity in WordNet. In: López de Mántaras, R., Saitta, L. (Eds.), *Proceedings of the 16th European Conference on Artificial Intelligence (ECAI)*. Vol. 16. IOS Press, Valencia, Spain, pp. 1089–1094.
- Solé-Ribalta, A., Sánchez, D., Batet, M., Serratosa, F., Jan. 2014. Towards the estimation of feature-based semantic similarity using multiple ontologies. *Knowledge-Based Systems* 55, 101–113.
- Song, W., Li, C. H., Park, S. C., Jul. 2009. Genetic algorithm for text clustering using ontology and evaluating the validity of various semantic similarity measures. *Expert Systems with Applications* 36 (5), 9095–9104.
- Stanchev, L., 2 Jun. 2014. Creating a Similarity Graph from WordNet. In: *Proceedings of the 4th International Conference on Web Intelligence, Mining and Semantics (WIMS'14)*. Article No. 36. ACM.
- Strube, M., Ponzetto, S. P., 2006. WikiRelate! Computing semantic relatedness using Wikipedia. In: *Proceedings of the AAAI Conference*. Vol. 6. AAAI, pp. 1419–1424.
- Suzuki, J., Nagata, M., Jul. 2015. A Unified Learning Framework of Skip-Grams and Global Vectors. In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. ACL, Beijing, China, pp. 186–191.
- Tversky, A., Jul. 1977. Features of similarity. *Psychological Review* 84 (4), 327–352.
- Wang, Y., Zhou, Z., 20 Sep. 2009. Domain Ontology Generation Based on WordNet and Internet. In: *Proceedings of the International Conference on Management and Service Science, 2009. MASS '09*. IEEE, Wuhan, China, pp. 1–5.
- Wolke, A., Bichler, M., Chirigati, F., Steeves, V., Jul. 2016. Reproducible experiments on dynamic resource allocation in cloud data centers. *Information Systems* 59, 98–101.
- Wolke, A., Tsend-Ayush, B., Pfeiffer, C., Bichler, M., Aug. 2015. More than bin packing: Dynamic resource allocation strategies in cloud data centers. *Information Systems* 52, 83–95.
- Wu, Z., Palmer, M., 1994. Verbs Semantics and Lexical Selection. In: *Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics*. ACL '94. ACL, Stroudsburg, PA, USA, pp. 133–138.
- Yazdani, M., Popescu-Belis, A., Jan. 2013. Computing text semantic relatedness using the contents and links of a hypertext encyclopedia. *Artificial Intelligence* 194, 176–202.
- Yeh, E., Ramage, D., Manning, C. D., Agirre, E., Soroa, A., 2009. WikiWalk: Random Walks on Wikipedia for Semantic Relatedness. In: *Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing. TextGraphs-4*. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 41–49.
- Yuan, Q., Yu, Z., Wang, K., Dec. 2013. A New Model of Information Content for Measuring the Semantic Similarity between Concepts. In: *Proceedings of the International Conference on Cloud Computing and Big Data (CloudCom-Asia 2013)*. IEEE Computer Society, pp. 141–146.
- Zhou, Z., Wang, Y., Gu, J., 2008a. A new model of information content for semantic similarity in WordNet. In: *Proc .of the Second International Conference on Future Generation Communication and Networking Symposia (FGCNS'08)*. Vol. 3. IEEE, pp. 85–89.
- Zhou, Z., Wang, Y., Gu, J., Nov. 2008b. New model of semantic similarity measuring in WordNet. In: *Proceedings of the 3rd International Conference on Intelligent System and Knowledge Engineering (ISKE 2008)*. Vol. 1. IEEE, pp. 256–261.

Average - r		Classic IC-based measures				Monotone trans. from classic measures				hybrid IC-based measures (shortest path length)				
IC models	Resnik	Lin	J&C	P&S	FaITH	Meng ₁₂	Garla	cosJ&C	Li _{s9}	Zhou	Meng ₁₄	Gao	coswJ&C	Avg
Seco et al. (2004)	0.7385	0.7731	0.7705	0.7784	0.7743	0.7758	0.7762	0.7790	0.7486	0.7859	0.7373	0.7241	0.7727	0.7642
Adhikari et al. (2015)	0.7373	0.7714	0.7645	0.7717	0.7794	0.7792	0.7818	0.7762	0.7528	0.7744	0.7219	0.7241	0.7784	0.7626
Meng et al. (2012)	0.7367	0.7704	0.7632	0.7708	0.7794	0.7790	0.7821	0.7758	0.7539	0.7732	0.7185	0.7241	0.7784	0.7620
Yuan et al. (2013)	0.7372	0.7744	0.7613	0.7745	0.7826	0.7830	0.7788	0.7736	0.7511	0.7738	0.7187	0.7241	0.7695	0.7617
Sánchez et al. (2011)	0.7459	0.7612	0.7681	0.7174	0.7780	0.7752	0.6842	0.7811	0.7674	0.7702	0.7058	0.7733	0.7836	0.7547
Sánchez and Batet (2012)	0.7424	0.7727	0.7668	0.7408	0.7735	0.7753	0.6223	0.7752	0.7725	0.7694	0.7355	0.7756	0.7711	0.7533
Blanchard et al. (2008)	0.7371	0.7708	0.7660	0.7392	0.7708	0.7727	0.6232	0.7752	0.7737	0.7690	0.7372	0.7760	0.7709	0.7525
CPRefHyponyms	0.7325	0.7682	0.7651	0.7357	0.7686	0.7704	0.6247	0.7738	0.7748	0.7681	0.7374	0.7762	0.7727	0.7514
Resnik (1995)	0.7416	0.7690	0.7632	0.7404	0.7736	0.7742	0.6195	0.7718	0.7714	0.7660	0.7290	0.7768	0.7717	0.7514
CPRefCosine	0.7297	0.7689	0.7655	0.7363	0.7683	0.7704	0.6292	0.7729	0.7736	0.7689	0.7365	0.7757	0.7659	0.7509
CPRefLeaves	0.7319	0.7678	0.7633	0.7338	0.7689	0.7705	0.6244	0.7726	0.7750	0.7662	0.7369	0.7762	0.7719	0.7507
CPHyponyms	0.7330	0.7673	0.7659	0.7332	0.7688	0.7705	0.6236	0.7739	0.7717	0.7686	0.7300	0.7742	0.7727	0.7502
CPRefCosineLeaves	0.7301	0.7683	0.7626	0.7331	0.7689	0.7706	0.6279	0.7710	0.7738	0.7659	0.7359	0.7759	0.7654	0.7499
Zhou et al. (2008a)	0.7236	0.7418	0.7411	0.7497	0.7734	0.7676	0.7792	0.7645	0.7669	0.7694	0.6753	0.7241	0.7677	0.7496
CPRefLeaSubRatio	0.7352	0.7693	0.7617	0.7322	0.7708	0.7723	0.6120	0.7691	0.7747	0.7636	0.7363	0.7764	0.7708	0.7496
CPRefCorpus	0.7364	0.7646	0.7561	0.7324	0.7678	0.7686	0.6376	0.7645	0.7749	0.7597	0.7328	0.7782	0.7704	0.7495
CPCosine	0.7340	0.7687	0.7676	0.7344	0.7694	0.7712	0.5819	0.7750	0.7660	0.7692	0.7335	0.7698	0.7708	0.7470
CPLeaves	0.7326	0.7670	0.7641	0.7314	0.7691	0.7705	0.5457	0.7726	0.7718	0.7667	0.7297	0.7743	0.7719	0.7436
CPRefLogistic	0.7148	0.7550	0.7530	0.7221	0.7670	0.7659	0.6278	0.7657	0.7720	0.7561	0.7324	0.7656	0.7646	0.7432
CPRefLogisticLeaves	0.7152	0.7548	0.7529	0.7217	0.7677	0.7664	0.6273	0.7657	0.7716	0.7560	0.7313	0.7652	0.7649	0.7431
CPCorpus	0.7353	0.7630	0.7553	0.7288	0.7675	0.7681	0.5600	0.7641	0.7715	0.7589	0.7247	0.7766	0.7704	0.7419
CPLogistic	0.7142	0.7659	0.7296	0.7125	0.7714	0.7727	0.5985	0.7477	0.7699	0.7335	0.7181	0.7716	0.7573	0.7356
CPRefUniform	0.6610	0.7633	0.6704	0.6671	0.7742	0.7737	0.5982	0.6876	0.7635	0.6745	0.7067	0.7677	0.7334	0.7109
CPUniform	0.6135	0.7031	0.6376	0.6037	0.7362	0.7297	0.5877	0.6557	0.7632	0.6422	0.6538	0.7624	0.7334	0.6786
Hadj Taieb et al. (2014a)	0.4233	0.6765	0.3267	0.4196	0.6882	0.6857	0.4890	0.3278	0.7481	0.3311	0.6616	0.7549	0.3278	0.5277
Best column value	0.7459	0.7744	0.7705	0.7784	0.7826	0.7830	0.7821	0.7811	0.7750	0.7859	0.7374	0.7782	0.7836	0.7642
Average per column*	0.7246	0.7633	0.7515	0.7309	0.7704	0.7706	0.6481	0.7627	0.7678	0.7571	0.7231	0.7628	0.7675	
Std. deviation*	0.0290	0.0147	0.0317	0.0355	0.0085	0.0096	0.0740	0.0292	0.0082	0.0322	0.0205	0.0207	0.0117	
Peak ratio*	0.7348	0.7538	0.6004	1.3370	1.4430	1.2956	1.8116	0.6297	0.8847	0.8956	0.6938	0.7420	1.3759	

Table 10: Average on all datasets of the Pearson (r) correlation values for each pair (IC model, IC measure). The IC models in bold are the new methods introduced in this work. The bold values represent the best score within each column, whilst the bold underlined value is the best overall score on all datasets for any combination of IC-based similarity measure with any IC model. Last row shows the best values within each column. The baseline is defined by the corpus-based Resnik IC model. The rows are arranged in descending order according to the average score in last column. (*) The Hadj Taieb et al. (2014a) IC model can be considered as an outlier, thus, we excluded it from the computation of these statistics. The peak ratio (pr) for a random variable X is defined by the equation $pr(X) = \frac{\max(X) - \bar{X}}{\sigma(X)}$, $\sigma(X)$ being the standard deviation of X .

Average - ρ		Classic IC-based measures				Monotone trans. from classic measures				hybrid IC-based measures (shortest path length)				
IC models	Resnik	Lin	J&C	P&S	FaITH	Meng12	Garla	cosJ&C	Li _{s9}	Zhou	Meng14	Gao	coswJ&C	Avg
Seco et al. (2004)	0.6987	0.7323	0.7375	0.7423	0.7323	0.7323	0.7375	0.7375	0.7231	0.7515	0.7423	0.7294	0.7416	0.7337
Yuan et al. (2013)	0.6997	0.7284	0.7363	0.7402	0.7284	0.7284	0.7363	0.7363	0.7242	0.7430	0.7407	0.7294	0.7395	0.7316
Zhou et al. (2008a)	0.6974	0.7250	0.7353	0.7267	0.7250	0.7250	0.7353	0.7353	0.7247	0.7459	0.7393	0.7294	0.7488	0.7302
Meng et al. (2012)	0.6961	0.7204	0.7387	0.7317	0.7204	0.7204	0.7387	0.7387	0.7194	0.7417	0.7366	0.7294	0.7541	0.7297
Adhikari et al. (2015)	0.6960	0.7206	0.7362	0.7292	0.7206	0.7206	0.7362	0.7362	0.7196	0.7417	0.7359	0.7294	0.7521	0.7288
Sánchez et al. (2011)	0.7008	0.7266	0.7402	0.6804	0.7266	0.7266	0.7402	0.7402	0.7128	0.7411	0.7411	0.7317	0.7579	0.7282
CPRefCosine	0.6963	0.7250	0.7297	0.7131	0.7250	0.7250	0.7297	0.7297	0.7238	0.7344	0.7419	0.7285	0.7394	0.7263
Sánchez and Batet (2012)	0.6957	0.7264	0.7281	0.7148	0.7264	0.7264	0.7281	0.7281	0.7196	0.7312	0.7386	0.7274	0.7367	0.7252
Blanchard et al. (2008)	0.6973	0.7242	0.7274	0.7131	0.7242	0.7242	0.7274	0.7274	0.7182	0.7340	0.7387	0.7278	0.7365	0.7247
Resnik (1995)	0.7005	0.7219	0.7303	0.7008	0.7219	0.7219	0.7303	0.7303	0.7167	0.7322	0.7437	0.7357	0.7344	0.7247
CPRefHyponyms	0.7007	0.7222	0.7255	0.7112	0.7222	0.7222	0.7255	0.7255	0.7211	0.7285	0.7399	0.7323	0.7403	0.7244
CPCosine	0.6923	0.7229	0.7300	0.7124	0.7229	0.7229	0.7283	0.7300	0.7132	0.7345	0.7370	0.7313	0.7384	0.7243
CPRefLeaves	0.6990	0.7242	0.7247	0.7104	0.7242	0.7242	0.7247	0.7247	0.7200	0.7325	0.7394	0.7269	0.7411	0.7243
CPHyponyms	0.6897	0.7202	0.7350	0.7128	0.7202	0.7202	0.7333	0.7350	0.7140	0.7385	0.7346	0.7212	0.7403	0.7242
CPRefCosineLeaves	0.6960	0.7210	0.7261	0.7114	0.7210	0.7210	0.7261	0.7261	0.7235	0.7320	0.7400	0.7311	0.7375	0.7241
CPLeaves	0.6882	0.7198	0.7262	0.7080	0.7198	0.7198	0.7245	0.7262	0.7159	0.7313	0.7341	0.7243	0.7411	0.7215
CPRefLeaSubRatio	0.7030	0.7209	0.7169	0.7028	0.7209	0.7209	0.7169	0.7169	0.7200	0.7213	0.7323	0.7306	0.7342	0.7198
CPLogistic	0.6938	0.7178	0.7189	0.6888	0.7178	0.7178	0.7190	0.7189	0.7173	0.7190	0.7328	0.7242	0.7189	0.7158
CPRefCorpus	0.6940	0.7144	0.7105	0.6923	0.7144	0.7144	0.7105	0.7105	0.7224	0.7160	0.7386	0.7356	0.7262	0.7154
CPCorpus	0.6893	0.7076	0.7134	0.6858	0.7076	0.7076	0.7111	0.7134	0.7166	0.7181	0.7385	0.7348	0.7262	0.7131
CPRefLogisticLeaves	0.6711	0.6940	0.7150	0.6740	0.6940	0.6940	0.7150	0.7150	0.7162	0.7198	0.7228	0.7078	0.7353	0.7057
CPRefLogistic	0.6714	0.6936	0.7107	0.6713	0.6936	0.6936	0.7107	0.7107	0.7153	0.7193	0.7232	0.7090	0.7318	0.7042
CPRefUniform	0.6604	0.7103	0.6806	0.6481	0.7103	0.7103	0.6806	0.6806	0.7133	0.6846	0.7299	0.7186	0.7324	0.6969
CPUniform	0.6166	0.6987	0.6818	0.6063	0.6987	0.6987	0.6844	0.6817	0.7152	0.6845	0.7135	0.7170	0.7324	0.6869
Hadj Taieb et al. (2014a)	0.5364	0.6570	0.4627	0.5164	0.6570	0.6570	0.4627	0.4627	0.7030	0.4685	0.6964	0.7136	0.4627	0.5735
Best column value	0.7030	0.7323	0.7402	0.7423	0.7323	0.7323	0.7402	0.7402	0.7247	0.7515	0.7437	0.7357	0.7579	0.7337
Average per column*	0.6893	0.7183	0.7231	0.7012	0.7183	0.7183	0.7229	0.7231	0.7186	0.7282	0.7356	0.7268	0.7382	0.7337
Std. deviation*	0.0193	0.0101	0.0158	0.0289	0.0101	0.0101	0.0154	0.0158	0.0036	0.0160	0.0072	0.0077	0.0092	0.0092
Peak ratio*	0.7066	1.3971	1.0843	1.4244	1.3967	1.3971	1.1229	1.0841	1.7198	1.4520	1.1188	1.1668	2.1322	0.7337

Table 11: Average on all datasets of the Spearman (ρ) correlation values for each pair (IC model, IC measure). The IC models in bold are the new methods introduced in this work. The bold values represent the best score within each column, whilst the bold underlined value is the best overall score on all datasets for any combination of IC-based similarity measure with any IC model. Last row shows the best values within each column. The baseline is defined by the corpus-based Resnik IC model. The rows are arranged in descending order according to the average score in last column. (*) The [Hadj Taieb et al. \(2014a\)](#) IC model can be considered as an outlier, thus, we excluded it from the computation of these statistics. The peak ratio (pr) for a random variable X is defined by the equation $pr(X) = \frac{\max(X) - \bar{X}}{\sigma(X)}$, $\sigma(X)$ being the standard deviation of X .

Sim. measures	IC model	Pearson (r)				Spearman (ρ)				Overall average scores				
		RG65	MC28	Agirre201	P&S _{full}	SimLex	RG65	MC28	Agirre201	P&S _{full}	SimLex	r	ρ	both
coswJ&C (L&G)	Sánchez (2011)	.8770	.8710	.6933	.8850	.5918	.8352	.8773	.6670	.8226	.5873	.7836	.7579	.7708
Zhou et al	Seco et al	.8728	.8541	.6839	.8949	.6237	.8245	.8466	.6619	.8144	.6101	.7859	.7515	.7687
cosJ&C (L&G)	Sánchez (2011)	.8752	.8476	.6890	.8996	.5941	.8034	.8492	.6576	.8003	.5906	.7811	.7402	.7607
Pirró & Seco	Seco et al	.8622	.8463	.6890	.8970	.5975	.8012	.8678	.6643	.7919	.5862	.7784	.7423	.7603
Taïeb et al	not apply	.8670	.8248	.7123	.9068	.6093	.7972	.8077	.6633	.7973	.5960	.7840	.7323	.7582
Gao et al	Resnik	.8709	.8336	.6731	.8909	.6152	.8153	.8082	.6469	.8089	.5994	.7768	.7357	.7563
Jiang & Conrath	Sánchez (2011)	.8619	.8595	.6591	.8762	.5838	.8034	.8492	.6576	.8003	.5906	.7681	.7402	.7542
Meng & Gu	Seco et al	.8596	.8144	.6969	.9031	.6048	.7972	.8314	.6530	.7911	.5888	.7758	.7323	.7540
P&S FaITH	Seco et al	.8565	.8094	.6966	.9042	.6046	.7972	.8314	.6530	.7911	.5888	.7743	.7323	.7533
Lin	Seco et al	.8609	.8240	.6850	.8945	.6010	.7972	.8314	.6530	.7911	.5888	.7731	.7323	.7527
Li _{s3} et al	not apply	.8625	.8355	.6675	.8853	.6059	.8106	.8144	.6313	.7986	.5918	.7713	.7294	.7504
Li _{s9} et al	Zhou et al	.8438	.8102	.6999	.8897	.5912	.7932	.7986	.6612	.7903	.5804	.7669	.7247	.7458
Li _{s4} et al	not apply	.8598	.8281	.6669	.8787	.6052	.7970	.7729	.6443	.7879	.5875	.7677	.7179	.7428
Sánchez et al	not apply	.8477	.8062	.6751	.8703	.5940	.7843	.7908	.6505	.7895	.5785	.7587	.7187	.7387
Meng et al	Resnik	.8434	.8064	.6006	.8286	.5659	.8227	.8135	.6663	.8105	.6057	.7290	.7437	.7363
Al-Mubaid	not apply	.8075	.7906	.6503	.8530	.5756	.8123	.8056	.6515	.8070	.5778	.7354	.7308	.7331
Resnik	CPReFLSRat	.8232	.7934	.6727	.8740	.5124	.7641	.8395	.6424	.7578	.5112	.7352	.7030	.7191
Pedersen et al	not apply	.7807	.7585	.6064	.8398	.5509	.8106	.8144	.6313	.7986	.5918	.7072	.7294	.7183
Leacock & Ch.	not apply	.7939	.7538	.5950	.7777	.5740	.8106	.8144	.6313	.7986	.5918	.6989	.7294	.7141
Garla & Brandt	Sánchez (2011)	.7690	.7198	.5733	.8470	.5117	.8034	.8492	.6576	.8003	.5906	.6842	.7402	.7122
Rada et al	not apply	.7708	.7294	.5798	.7506	.5651	.8106	.8144	.6313	.7986	.5918	.6791	.7294	.7043
Wu & Palmer	not apply	.7703	.7746	.6048	.7761	.5374	.7492	.7525	.6368	.7458	.5417	.6926	.6852	.6889
Best column values		.8770	.8710	.7123	.9068	.6237	.8352	.8773	.6670	.8226	.6101	.7859	.7579	.7708

Table 12: Results for each similarity measure on each dataset: Pearson (r) and Spearman (ρ) correlation coefficients, and averaged overall scores. Each IC-based similarity measure is evaluated with its best performing IC model in average, in accordance with the average Spearman correlation values in table 9. The bold values represent the best score within each column. The rows are arranged in descending order according to the overall average score in last column. The baseline is defined by the Jiang-Conrath similarity measure.

IC models	Resnik (1995)	Seco et al. (2004)	Blanchard et al. (2008)	Zhou et al. (2008a)	Sánchez et al. (2011)	Sánchez and Batet (2012)	Meng et al. (2012)	Yuan et al. (2013)	Hadj Taieb et al. (2014a)	Adhikari et al. (2015)	CondProbHyponyms	CondProbUniform	CondProbLeaves	CondProbCosine	CondProbLogistic	CondProbCorpus	CondProbRefHyponyms	CondProbRefUniform	CondProbRefLeaves	CondProbRefCosine	CondProbRefLogistic	CondProbRefCosineLeaves	CondProbRefLogisticLeaves	CPRefLeavesSubsumersRatio
Resnik (1995)	x	—	—	—	—	—	—	—	—	—	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Seco et al. (2004)	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Blanchard et al. (2008)	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Zhou et al. (2008a)	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Sánchez et al. (2011)	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Sánchez and Batet (2012)	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Meng et al. (2012)	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Yuan et al. (2013)	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Hadj Taieb et al. (2014a)	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Adhikari et al. (2015)	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
CondProbHyponyms	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
CondProbUniform	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
CondProbLeaves	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
CondProbCosine	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
CondProbLogistic	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
CondProbCorpus	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
CondProbRefHyponyms	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
CondProbRefUniform	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
CondProbRefLeaves	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
CondProbRefCosine	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
CondProbRefLogistic	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
CondProbRefCorpus	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
CondProbRefCosineLeaves	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
CondProbRefLogisticLeaves	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
CPRefLeavesSubsumersRatio	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x

Table 13: Summary of the statistical significance analysis between IC models derived from the pairwise raw p-values shown in table 25. The p-values are computed using a one-sided t-student distribution for the paired vectors of Spearman correlation values reported by each pair of IC models on each dataset. Each row shows a 'x' whenever the row IC model obtains a statistically significant higher performance than the column IC model. Thus, the rows show the IC models that are outperformed by each IC model on the left, whilst the columns show the IC models that outperform each IC model on the top. For instance, the Seco et al. (2004) IC model outperforms all IC models with the only exception of the Sánchez et al. (2011) IC model. On the other hand, a first glance to the columns shows that the Seco et al. (2004) and Sánchez et al. (2011) IC models are the only ones that are not outperformed by another IC model. Bold values in uppercase 'X' show the outperformance of the refined IC models as regard their corresponding non-refined models proving the main hypothesis.

Similarity measures	Lastra & García (2015), coswJ&C	Zhou et al. (2008b)	Lastra & García (2015), cosJ&C	Pirró and Seco (2008)	Hadj Taieb et al. (2014b)	Gao et al. (2015)	Jiang and Conrath (1997)	Meng and Gu (2012)	Pirró and Euzenat (2010)	Lim (1998)	Li et al. (2003), strat. 3	Li et al. (2003), strat. 9	Li et al. (2003), strat. 4	Sánchez et al. (2012)	Meng et al. (2014)	Al-Mubaid and Nguyen (2009)	Resnik (1995)	Pedersen et al. (2007)	Leacock and Chodorow (1998)	Garla and Brandt (2012)	Rada et al. (1989)	Wu and Palmer (1994)
Lastra & García (2015), coswJ&C	—	—	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Zhou et al. (2008b)	—	—	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Lastra & García (2015), cosJ&C	—	—	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Pirró and Seco (2008)	—	—	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Hadj Taieb et al. (2014b)	—	—	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Gao et al. (2015)	—	—	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Jiang and Conrath (1997)	—	—	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Meng and Gu (2012)	—	—	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Pirró and Euzenat (2010)	—	—	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Lin (1998)	—	—	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Li et al. (2003), strat. 3	—	—	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Li et al. (2003), strat. 9	—	—	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Li et al. (2003), strat. 4	—	—	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Sánchez et al. (2012)	—	—	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Meng et al. (2014)	—	—	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Al-Mubaid and Nguyen (2009)	—	—	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Resnik (1995)	—	—	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Pedersen et al. (2007)	—	—	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Leacock and Chodorow (1998)	—	—	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Garla and Brandt (2012)	—	—	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Rada et al. (1989)	—	—	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Wu and Palmer (1994)	—	—	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x

Table 14: Summary of the statistical significance analysis between similarity measures derived from the pairwise raw p-values shown in table 26. The p-values are computed using a one-sided t-student distribution for the paired vectors of Spearman correlation values reported by each pair of similarity measures on each dataset. Each row shows a 'x' whenever the row measure obtains a statistically significant higher performance than the respective column measure. Thus, the rows show the similarity measures that are outperformed by each measure on the left, whilst the columns show the similarity measures that outperform the measures on the top. For instance, the coswJ&C and Zhou et al. (2008b) similarity measures outperform most similarity measures, whilst a first glance to the columns show that the Wu and Palmer (1994) measure is outperformed by the remaining similarity measures.

RG65 - r	Classic IC-based measures				Monotone trans. from classic measures				hybrid IC-based measures (shortest path length)				
	Resnik	Lin	J&C	P&S	FaITH	Meng12	Garla	cosJ&C	Li9	Zhou	Meng14	Gao	coswJ&C
Resnik (1995)	0.8345	0.8589	0.8561	0.8335	0.8585	0.8609	0.6950	0.8653	0.8617	0.8582	0.8434	0.8709	0.8703
Seco et al. (2004)	0.8326	0.8609	0.8546	0.8622	0.8565	0.8596	0.8571	0.8642	0.8241	0.8728	0.8486	0.7992	0.8634
Blanchard et al. (2008)	0.8310	0.8589	0.8493	0.8303	0.8536	0.8571	0.6957	0.8607	0.8618	0.8525	0.8482	0.8669	0.8614
Zhou et al. (2008a)	0.8080	0.8259	0.8286	0.8334	0.8589	0.8539	0.8662	0.8558	0.8438	0.8574	0.7749	0.7992	0.8610
Sánchez et al. (2011)	0.8409	0.8530	0.8619	0.8105	0.8683	0.8663	0.7690	0.8752	0.8586	0.8639	0.8147	0.8682	0.8770
Sánchez and Batet (2012)	0.8355	0.8616	0.8508	0.8332	0.8565	0.8600	0.6940	0.8606	0.8615	0.8535	0.8475	0.8678	0.8614
Meng et al. (2012)	0.8260	0.8608	0.8598	0.8586	0.8658	0.8670	0.8717	0.8723	0.8282	0.8638	0.8285	0.7992	0.8747
Yuan et al. (2013)	0.8243	0.8621	0.8505	0.8607	0.8649	0.8675	0.8609	0.8632	0.8231	0.8629	0.8273	0.7992	0.8624
Hadj Taieb et al. (2014a)	0.4658	0.7825	0.4543	0.5090	0.7933	0.7924	0.5939	0.4552	0.8444	0.4586	0.7693	0.8527	0.4552
Adhikari et al. (2015)	0.8264	0.8612	0.8609	0.8592	0.8650	0.8664	0.8707	0.8722	0.8267	0.8650	0.8321	0.7992	0.8747
IC models introduced by Lastra-Díaz and García-Serrano (2015a)													
CPHyponyms	0.8283	0.8587	0.8562	0.8286	0.8556	0.8585	0.7033	0.8658	0.8606	0.8586	0.8426	0.8665	0.8645
CPUniform	0.6844	0.7958	0.7228	0.6840	0.8425	0.8323	0.6721	0.7473	0.8549	0.7277	0.7520	0.8569	0.8331
CPLeaves	0.8272	0.8575	0.8535	0.8263	0.8550	0.8578	0.7009	0.8635	0.8605	0.8560	0.8418	0.8662	0.8625
CPCosine	0.8294	0.8591	0.8540	0.8265	0.8550	0.8581	0.6609	0.8634	0.8574	0.8555	0.8456	0.8658	0.8624
CPLogistic	0.8021	0.8681	0.8260	0.8147	0.8652	0.8692	0.6899	0.8501	0.8605	0.8300	0.8363	0.8656	0.8595
CPCorpus	0.8296	0.8560	0.8527	0.8223	0.8575	0.8591	0.7275	0.8633	0.8618	0.8552	0.8405	0.8706	0.8682
IC models introduced in this work													
CPRefHyponyms	0.8242	0.8547	0.8497	0.8256	0.8514	0.8543	0.7001	0.8610	0.8615	0.8527	0.8463	0.8648	0.8645
CPRefUniform	0.7253	0.8540	0.7484	0.7453	0.8651	0.8652	0.6810	0.7703	0.8551	0.7530	0.8103	0.8615	0.8331
CPRefLeaves	0.8230	0.8533	0.8469	0.8231	0.8508	0.8535	0.6980	0.8587	0.8614	0.8499	0.8454	0.8644	0.8625
CPRefCosine	0.8231	0.8573	0.8506	0.8252	0.8532	0.8563	0.7115	0.8609	0.8617	0.8539	0.8467	0.8663	0.8621
CPRefLogistic	0.7991	0.8440	0.8421	0.8125	0.8536	0.8535	0.7173	0.8597	0.8502	0.8454	0.8355	0.8444	0.8648
CPRefCorpus	0.8264	0.8528	0.8484	0.8210	0.8532	0.8550	0.7217	0.8588	0.8627	0.8512	0.8448	0.8684	0.8682
CPRefCosineLeaves	0.8221	0.8546	0.8457	0.8210	0.8517	0.8544	0.7044	0.8568	0.8613	0.8491	0.8451	0.8656	0.8587
CPRefLogisticLeaves	0.7991	0.8430	0.8427	0.8120	0.8535	0.8532	0.7135	0.8599	0.8495	0.8458	0.8336	0.8438	0.8653
CPRefLeavesSubsumersRatio	0.8232	0.8564	0.8543	0.8292	0.8535	0.8564	0.6850	0.8621	0.8614	0.8560	0.8459	0.8649	0.8682
Best column value	0.8409	0.8681	0.8619	0.8622	0.8683	0.8692	0.8717	0.8752	0.8627	0.8728	0.8486	0.8709	0.8770

Table 15: Pearson (r) correlation coefficients for all the IC models and measures in the RG65 dataset using WordNet 3.0. The bold values represent the best score within each column, whilst the bold underlined value is the best overall score in the dataset for any combination of IC-based similarity measure with any IC model. Last row shows the best values within each column.

RG65 - ρ	Classic IC-based measures				Monotone trans. from classic measures				hybrid IC-based measures (shortest path length)				
	Resnik	Lin	J&C	P&S	FaITH	Meng12	Garla	cosJ&C	Li19	Zhou	Meng14	Gao	coswJ&C
Resnik (1995)	0.7777	0.7761	0.7831	0.7633	0.7761	0.7761	0.7831	0.7831	0.7868	0.7848	0.8227	0.8153	0.8009
Seco et al. (2004)	0.7735	0.7972	0.7866	0.8012	0.7972	0.7850	0.7866	0.7866	0.7939	0.8245	0.8204	0.8106	0.8061
Blanchard et al. (2008)	0.7602	0.7850	0.7788	0.7816	0.7850	0.7850	0.7788	0.7788	0.7867	0.7855	0.8174	0.7980	0.8014
Zhou et al. (2008a)	0.7690	0.7881	0.8051	0.7958	0.7881	0.7881	0.8051	0.8051	0.7932	0.8244	0.8147	0.8106	0.8244
Sánchez et al. (2011)	0.7714	0.7944	0.8034	0.7671	0.7944	0.7944	0.8034	0.8034	0.7846	0.8081	0.8201	0.8108	0.8352
Sánchez and Batet (2012)	0.7706	0.7911	0.7779	0.7856	0.7911	0.7911	0.7779	0.7779	0.7869	0.7854	0.8170	0.8000	0.8003
Meng et al. (2012)	0.7607	0.7817	0.8166	0.8140	0.7817	0.7817	0.8166	0.8166	0.7886	0.8177	0.8131	0.8106	0.8408
Yuan et al. (2013)	0.7742	0.7919	0.8050	0.8206	0.7919	0.7919	0.8050	0.8050	0.7932	0.8195	0.8151	0.8106	0.8174
Hadj Taieb et al. (2014a)	0.5990	0.7417	0.6126	0.6123	0.7417	0.7417	0.6126	0.6126	0.7719	0.6182	0.7817	0.7787	0.6126
Adhikari et al. (2015)	0.7609	0.7829	0.8086	0.8084	0.7829	0.7829	0.8086	0.8086	0.7886	0.8178	0.8114	0.8106	0.8369
IC models introduced by Lastra-Díaz and García-Serrano (2015a)													
CPHyponyms	0.7561	0.7888	0.8017	0.7941	0.7888	0.7888	0.8017	0.8017	0.7826	0.8052	0.8154	0.7959	0.8071
CPUniform	0.7003	0.7786	0.7613	0.6930	0.7786	0.7786	0.7613	0.7613	0.7888	0.7638	0.7928	0.7880	0.8146
CPLeaves	0.7549	0.7868	0.7877	0.7848	0.7868	0.7868	0.7877	0.7877	0.7825	0.7951	0.8143	0.7965	0.8081
CPCosine	0.7710	0.7896	0.7835	0.7830	0.7896	0.7896	0.7835	0.7835	0.7874	0.7882	0.8174	0.8116	0.8037
CPLogistic	0.7833	0.7966	0.7993	0.7792	0.7966	0.7966	0.7993	0.7993	0.7899	0.8003	0.8145	0.7984	0.7977
CPCorpus	0.7702	0.7652	0.7722	0.7515	0.7652	0.7652	0.7722	0.7722	0.7868	0.7786	0.8208	0.8148	0.7833
IC models introduced in this work													
CPRefHyponyms	0.7651	0.7786	0.7765	0.7791	0.7786	0.7786	0.7765	0.7765	0.7882	0.7812	0.8142	0.8018	0.8071
CPRefUniform	0.7273	0.7773	0.7475	0.7220	0.7773	0.7773	0.7475	0.7475	0.7826	0.7526	0.8039	0.7841	0.8146
CPRefLeaves	0.7627	0.7804	0.7760	0.7779	0.7804	0.7804	0.7760	0.7760	0.7875	0.7861	0.8125	0.7880	0.8081
CPRefCosine	0.7606	0.7893	0.7835	0.7794	0.7893	0.7893	0.7835	0.7835	0.7911	0.7878	0.8197	0.7996	0.8137
CPRefLogistic	0.7348	0.7404	0.7645	0.7320	0.7404	0.7404	0.7645	0.7645	0.7570	0.7756	0.7782	0.7492	0.8044
CPRefCorpus	0.7612	0.7635	0.7580	0.7482	0.7635	0.7635	0.7580	0.7580	0.7897	0.7649	0.8131	0.8021	0.7833
CPRefCosineLeaves	0.7600	0.7796	0.7752	0.7753	0.7796	0.7796	0.7752	0.7752	0.7920	0.7825	0.8158	0.8013	0.8061
CPRefLogisticLeaves	0.7347	0.7407	0.7732	0.7351	0.7407	0.7407	0.7732	0.7732	0.7587	0.7785	0.7772	0.7485	0.8112
CPRefLeavesSubsumersRatio	0.7641	0.7728	0.7810	0.7724	0.7728	0.7728	0.7810	0.7810	0.7877	0.7867	0.8090	0.7965	0.8129
Best column value	0.7833	0.7972	0.8166	0.8206	0.7972	0.7972	0.8166	0.8166	0.7939	0.8245	0.8227	0.8153	0.8408

Table 16: Spearman (ρ) correlation coefficients for all the IC models and measures in the RG65 dataset using WordNet 3.0. The bold values represent the best score within each column, whilst the bold underlined value is the best overall score in the dataset for any combination of IC-based similarity measure with any IC model. Last row shows the best values within each column.

MC28 - r	Classic IC-based measures				Monotone trans. from classic measures				hybrid IC-based measures (shortest path length)				
	Resnik	Lin	J&C	P&S	FaITH	Meng12	Garla	cosJ&C	Li _{s9}	Zhou	Meng14	Gao	coswJ&C
Resnik (1995)	0.7929	0.8350	0.8809	0.8466	0.8233	0.8276	0.6611	0.8710	0.8260	0.8795	0.8064	0.8336	0.8789
Seco et al. (2004)	0.7834	0.8240	0.8557	0.8463	0.8094	0.8144	0.8299	0.8427	0.7730	0.8541	0.8107	0.7775	0.8369
Blanchard et al. (2008)	0.7823	0.8228	0.8528	0.8169	0.8060	0.8118	0.6694	0.8385	0.8267	0.8535	0.8122	0.8299	0.8356
Zhou et al. (2008a)	0.8179	0.8222	0.8282	0.8281	0.8344	0.8338	0.8410	0.8403	0.8102	0.8396	0.7638	0.7775	0.8492
Sánchez et al. (2011)	0.8189	0.8357	0.8595	0.8056	0.8343	0.8359	0.7198	0.8476	0.8236	0.8582	0.7912	0.8366	0.8710
Sánchez and Batet (2012)	0.7905	0.8261	0.8507	0.8177	0.8103	0.8159	0.6686	0.8411	0.8247	0.8511	0.8101	0.8282	0.8384
Meng et al. (2012)	0.8202	0.8393	0.8314	0.8330	0.8330	0.8361	0.8376	0.8393	0.7959	0.8350	0.8005	0.7775	0.8569
Yuan et al. (2013)	0.8084	0.8341	0.8347	0.8384	0.8277	0.8315	0.8320	0.8407	0.7881	0.8381	0.7953	0.7775	0.8350
Hadj Taieb et al. (2014a)	0.5391	0.6842	0.4587	0.4845	0.6899	0.6875	0.5516	0.4594	0.8049	0.4622	0.6792	0.8243	0.4594
Adhikari et al. (2015)	0.8211	0.8410	0.8331	0.8351	0.8331	0.8365	0.8366	0.8392	0.7949	0.8368	0.8040	0.7775	0.8552
IC models introduced by Lastra-Díaz and García-Serrano (2015a)													
CPHyponyms	0.7860	0.8215	0.8552	0.8110	0.8070	0.8124	0.6687	0.8429	0.8242	0.8548	0.8030	0.8280	0.8406
CPUniform	0.7408	0.7753	0.6483	0.6365	0.8039	0.7971	0.6250	0.6698	0.8190	0.6530	0.7276	0.8325	0.7925
CPLeaves	0.7869	0.8219	0.8511	0.8073	0.8080	0.8131	0.6694	0.8402	0.8248	0.8507	0.8039	0.8287	0.8387
CPCosine	0.7797	0.8208	0.8562	0.8127	0.8054	0.8109	0.6342	0.8445	0.8211	0.8563	0.8058	0.8316	0.8410
CPLogistic	0.7700	0.8142	0.8020	0.7687	0.8079	0.8119	0.6624	0.8095	0.8225	0.8046	0.7801	0.8249	0.8232
CPCorpus	0.7912	0.8290	0.8678	0.8281	0.8163	0.8209	0.6839	0.8613	0.8262	0.8668	0.8014	0.8338	0.8745
IC models introduced in this work													
CPRefHyponyms	0.7884	0.8242	0.8519	0.8134	0.8061	0.8122	0.6670	0.8412	0.8315	0.8522	0.8170	0.8348	0.8406
CPRefUniform	0.7317	0.8257	0.6998	0.7000	0.8223	0.8254	0.6487	0.7206	0.8193	0.7035	0.7782	0.8344	0.7925
CPRefLeaves	0.7891	0.8244	0.8476	0.8092	0.8073	0.8131	0.6675	0.8386	0.8322	0.8480	0.8177	0.8355	0.8387
CPRefCosine	0.7767	0.8215	0.8533	0.8144	0.8029	0.8093	0.6675	0.8455	0.8278	0.8540	0.8123	0.8309	0.8393
CPRefLogistic	0.7848	0.8226	0.8376	0.7999	0.8183	0.8210	0.6818	0.8374	0.8374	0.8396	0.8248	0.8288	0.8437
CPRefCorpus	0.7950	0.8312	0.8675	0.8324	0.8155	0.8206	0.6758	0.8597	0.8325	0.8670	0.8165	0.8383	0.8745
CPRefCosineLeaves	0.7805	0.8220	0.8456	0.8069	0.8046	0.8106	0.6680	0.8391	0.8288	0.8464	0.8141	0.8326	0.8365
CPRefLogisticLeaves	0.7881	0.8240	0.8383	0.8004	0.8202	0.8227	0.6836	0.8385	0.8374	0.8402	0.8252	0.8292	0.8444
CPRefLeavesSubsumersRatio	0.7934	0.8243	0.8363	0.7975	0.8115	0.8160	0.6565	0.8356	0.8315	0.8362	0.8143	0.8352	0.8417
Best column value	0.8211	0.8410	0.8809	0.8466	0.8344	0.8365	0.8410	0.8710	0.8374	0.8795	0.8252	0.8383	0.8789

Table 17: Pearson (r) correlation coefficients for all the IC models and measures in the MC28 dataset using WordNet 3.0. The bold values represent the best score within each column, whilst the bold underlined value is the best overall score in the dataset for any combination of IC-based similarity measure with any IC model. Last row shows the best values within each column.

MC28 - ρ IC models	Classic IC-based measures				Monotone trans. from classic measures				hybrid IC-based measures (shortest path length)				
	Resnik	Lin	J&C	P&S	FaITH	Meng12	Garla	cosJ&C	Li19	Zhou	Meng14	Gao	coswJ&C
Resnik (1995)	0.7935	0.8403	0.8882	0.8255	0.8403	0.8403	0.8882	0.8882	0.7734	0.8871	0.8135	0.8082	0.8923
Seco et al. (2004)	0.7947	0.8314	0.8727	0.8678	0.8314	0.8314	0.8727	0.8727	0.7956	0.8466	0.8053	0.8144	0.8801
Blanchard et al. (2008)	0.8239	0.8225	0.8518	0.8335	0.8225	0.8225	0.8518	0.8518	0.7824	0.8643	0.7973	0.8112	0.8749
Zhou et al. (2008a)	0.7971	0.8072	0.8244	0.8091	0.8072	0.8072	0.8244	0.8244	0.7986	0.8192	0.8042	0.8144	0.8600
Sánchez et al. (2011)	0.7937	0.8114	0.8492	0.7348	0.8114	0.8114	0.8492	0.8492	0.7635	0.8413	0.8058	0.7983	0.8773
Sánchez and Batet (2012)	0.7774	0.8212	0.8551	0.8329	0.8212	0.8212	0.8551	0.8551	0.7865	0.8518	0.7968	0.7972	0.8734
Meng et al. (2012)	0.8296	0.8080	0.8198	0.8050	0.8080	0.8080	0.8198	0.8198	0.7926	0.8198	0.8042	0.8144	0.8543
Yuan et al. (2013)	0.7971	0.8042	0.8274	0.8083	0.8042	0.8042	0.8274	0.8274	0.7975	0.8209	0.8031	0.8144	0.8285
Hadj Taieb et al. (2014a)	0.6340	0.6961	0.6102	0.4673	0.6961	0.6961	0.6102	0.6102	0.7882	0.6110	0.7429	0.8010	0.6102
Adhikari et al. (2015)	0.8296	0.8080	0.8198	0.8050	0.8080	0.8080	0.8198	0.8198	0.7926	0.8198	0.8036	0.8144	0.8526
IC models introduced by Lastra-Díaz and García-Serrano (2015a)													
CPHyponyms	0.8131	0.8034	0.8554	0.8187	0.8034	0.8034	0.8554	0.8554	0.7734	0.8589	0.7875	0.7842	0.8753
CPUniform	0.7564	0.7749	0.7281	0.6155	0.7749	0.7749	0.7281	0.7281	0.7761	0.7339	0.7746	0.7813	0.8077
CPLeaves	0.8073	0.8039	0.8389	0.8110	0.8039	0.8039	0.8389	0.8389	0.7824	0.8436	0.7875	0.7979	0.8741
CPCosine	0.7791	0.8116	0.8606	0.8324	0.8116	0.8116	0.8606	0.8606	0.7635	0.8688	0.7919	0.8012	0.8751
CPLogistic	0.7909	0.7752	0.8034	0.7377	0.7752	0.7752	0.8034	0.8034	0.7770	0.7968	0.7848	0.7957	0.8097
CPCorpus	0.7707	0.8058	0.8502	0.7946	0.8058	0.8058	0.8502	0.8502	0.7734	0.8537	0.7984	0.8075	0.8797
IC models introduced in this work													
CPRefHyponyms	0.8433	0.8313	0.8504	0.8274	0.8313	0.8313	0.8504	0.8504	0.7958	0.8507	0.8116	0.8276	0.8753
CPRefUniform	0.7738	0.7713	0.7287	0.6597	0.7713	0.7713	0.7287	0.7287	0.7729	0.7344	0.7837	0.7868	0.8077
CPRefLeaves	0.8376	0.8362	0.8409	0.8242	0.8362	0.8362	0.8409	0.8409	0.7912	0.8573	0.8116	0.8276	0.8741
CPRefCosine	0.8288	0.8217	0.8510	0.8296	0.8217	0.8217	0.8510	0.8510	0.8049	0.8600	0.8067	0.8134	0.8693
CPRefLogistic	0.7737	0.7757	0.8124	0.7547	0.7757	0.7757	0.8124	0.8124	0.8253	0.8280	0.8108	0.8135	0.8461
CPRefCorpus	0.8124	0.8368	0.8592	0.8181	0.8368	0.8368	0.8592	0.8592	0.7958	0.8649	0.8110	0.8333	0.8797
CPRefCosineLeaves	0.8310	0.8217	0.8466	0.8296	0.8217	0.8217	0.8466	0.8466	0.7999	0.8567	0.8077	0.8221	0.8655
CPRefLogisticLeaves	0.7737	0.7757	0.8138	0.7596	0.7757	0.7757	0.8138	0.8138	0.8258	0.8239	0.8119	0.8086	0.8452
CPRefLeavesSubsumersRatio	0.8395	0.8395	0.7960	0.7968	0.8395	0.8395	0.7960	0.7960	0.7895	0.8039	0.7905	0.8248	0.8354
Best column value	0.8433	0.8403	0.8882	0.8678	0.8403	0.8403	0.8882	0.8882	0.8258	0.8871	0.8135	0.8333	0.8923

Table 18: Spearman (ρ) correlation coefficients for all the IC models and measures in the MC28 dataset using WordNet 3.0. The bold values represent the best score within each column, whilst the bold underlined value is the best overall score in the dataset for any combination of IC-based similarity measure with any IC model. Last row shows the best values within each column.

Agirre201 - r IC models	Classic IC-based measures				Monotone trans. from classic measures				hybrid IC-based measures (shortest path length)				
	Resnik	Lin	J&C	P&S	FaITH	Meng12	Garla	cosJ&C	Li-s9	Zhou	Meng14	Gao	coswJ&C
Resnik (1995)	0.6716	0.6697	0.6349	0.6401	0.6913	0.6882	0.5007	0.6524	0.6726	0.6393	0.6006	0.6731	0.6490
Seco et al. (2004)	0.6629	0.6850	0.6724	0.6890	0.6966	0.6969	0.6885	0.6904	0.6856	0.6839	0.6185	0.6252	0.6784
Blanchard et al. (2008)	0.6615	0.6834	0.6716	0.6590	0.6939	0.6945	0.5046	0.6891	0.6776	0.6742	0.6184	0.6751	0.6803
Zhou et al. (2008a)	0.6453	0.6409	0.6288	0.6503	0.6848	0.6753	0.6815	0.6564	0.6999	0.6616	0.5591	0.6252	0.6638
Sánchez et al. (2011)	0.6592	0.6643	0.6591	0.6339	0.6946	0.6889	0.5733	0.6890	0.6676	0.6616	0.5842	0.6664	0.6933
Sánchez and Batet (2012)	0.6678	0.6848	0.6720	0.6602	0.6972	0.6973	0.5052	0.6878	0.6762	0.6743	0.6167	0.6747	0.6788
Meng et al. (2012)	0.6756	0.6817	0.6593	0.6863	0.7039	0.7008	0.6890	0.6747	0.6980	0.6719	0.6006	0.6252	0.6782
Yuan et al. (2013)	0.6760	0.6858	0.6543	0.6863	0.7061	0.7040	0.6834	0.6695	0.6944	0.6681	0.6046	0.6252	0.6633
Hadj Taieb et al. (2014a)	0.4165	0.6345	0.0790	0.3150	0.6490	0.6467	0.2674	0.0793	0.6827	0.0844	0.5885	0.6774	0.0793
Adhikari et al. (2015)	0.6755	0.6824	0.6606	0.6866	0.7041	0.7013	0.6890	0.6756	0.6975	0.6728	0.6028	0.6252	0.6786
IC models introduced by Lastra-Díaz and García-Serrano (2015a)													
CPHyponyms	0.6557	0.6751	0.6585	0.6470	0.6868	0.6874	0.5003	0.6748	0.6753	0.6618	0.6097	0.6723	0.6752
CPUniform	0.5713	0.6156	0.5597	0.5682	0.6516	0.6432	0.4932	0.5719	0.6617	0.5636	0.5505	0.6535	0.6233
CPLeaves	0.6569	0.6765	0.6601	0.6478	0.6888	0.6891	0.1151	0.6764	0.6755	0.6631	0.6096	0.6725	0.6773
CPCosine	0.6579	0.6789	0.6660	0.6521	0.6893	0.6902	0.4471	0.6805	0.6660	0.6676	0.6144	0.6598	0.6735
CPLogistic	0.6501	0.6618	0.6206	0.6290	0.6809	0.6796	0.4783	0.6380	0.6716	0.6242	0.5959	0.6653	0.6416
CPCorpus	0.6638	0.6603	0.6285	0.6326	0.6807	0.6779	0.5162	0.6426	0.6727	0.6339	0.5939	0.6728	0.6526
IC models introduced in this work													
CPRefHyponyms	0.6604	0.6817	0.6646	0.6557	0.6915	0.6924	0.5025	0.6798	0.6787	0.6675	0.6178	0.6763	0.6752
CPRefUniform	0.6533	0.6926	0.5964	0.6423	0.7137	0.7110	0.4780	0.6086	0.6621	0.6002	0.6027	0.6618	0.6233
CPRefLeaves	0.6612	0.6827	0.6658	0.6563	0.6931	0.6938	0.5040	0.6813	0.6790	0.6686	0.6177	0.6764	0.6773
CPRefCosine	0.6586	0.6798	0.6614	0.6550	0.6884	0.6898	0.4998	0.6736	0.6772	0.6647	0.6172	0.6763	0.6591
CPRefLogistic	0.6558	0.6735	0.6543	0.6488	0.6926	0.6907	0.4904	0.6695	0.6857	0.6570	0.6148	0.6794	0.6641
CPRefCorpus	0.6707	0.6682	0.6348	0.6416	0.6864	0.6843	0.5150	0.6484	0.6775	0.6398	0.6024	0.6772	0.6526
CPRefCosineLeaves	0.6608	0.6819	0.6638	0.6564	0.6917	0.6927	0.5036	0.6771	0.6776	0.6670	0.6169	0.6766	0.6637
CPRefLogisticLeaves	0.6565	0.6742	0.6535	0.6488	0.6943	0.6921	0.4920	0.6688	0.6858	0.6561	0.6140	0.6793	0.6643
CPRefLeavesSubsumersRatio	0.6727	0.6877	0.6664	0.6603	0.6977	0.6984	0.4949	0.6757	0.6786	0.6683	0.6191	0.6767	0.6709
Best column value	0.6760	0.6926	0.6724	0.6890	0.7137	0.7110	0.6890	0.6904	0.6999	0.6839	0.6191	0.6794	0.6933

Table 19: Pearson (r) correlation coefficients for all the IC models and measures in the Agirre201 dataset using WordNet 3.0. The bold values represent the best score within each column, whilst the bold underlined value is the best overall score in the dataset for any combination of IC-based similarity measure with any IC model. Last row shows the best values within each column.

Agirre201 - ρ IC models	Classic IC-based measures				Monotone trans. from classic measures				hybrid IC-based measures (shortest path length)				
	Resnik	Lin	J&C	P&S	FaITH	Meng12	Garla	cosJ&C	Li-s9	Zhou	Meng14	Gao	coswJ&C
Resnik (1995)	0.6320	0.6461	0.6382	0.6282	0.6461	0.6461	0.6382	0.6382	0.6503	0.6412	0.6663	0.6469	0.6331
Seco et al. (2004)	0.6308	0.6530	0.6612	0.6643	0.6530	0.6530	0.6612	0.6612	0.6560	0.6619	0.6606	0.6313	0.6482
Blanchard et al. (2008)	0.6309	0.6505	0.6577	0.6409	0.6505	0.6505	0.6577	0.6577	0.6536	0.6613	0.6594	0.6445	0.6465
Zhou et al. (2008a)	0.6460	0.6591	0.6524	0.6581	0.6591	0.6591	0.6524	0.6524	0.6612	0.6556	0.6631	0.6313	0.6591
Sánchez et al. (2011)	0.6408	0.6534	0.6576	0.6224	0.6534	0.6534	0.6576	0.6576	0.6482	0.6598	0.6601	0.6496	0.6670
Sánchez and Batet (2012)	0.6345	0.6494	0.6590	0.6399	0.6494	0.6494	0.6590	0.6590	0.6545	0.6590	0.6606	0.6476	0.6490
Meng et al. (2012)	0.6462	0.6581	0.6488	0.6560	0.6581	0.6581	0.6488	0.6488	0.6646	0.6532	0.6620	0.6313	0.6562
Yuan et al. (2013)	0.6481	0.6652	0.6505	0.6656	0.6652	0.6652	0.6505	0.6505	0.6656	0.6495	0.6678	0.6313	0.6466
Hadj Taieb et al. (2014a)	0.5343	0.6175	0.1556	0.4676	0.6175	0.6175	0.1556	0.1556	0.6600	0.1672	0.6310	0.6662	0.1556
Adhikari et al. (2015)	0.6457	0.6584	0.6497	0.6563	0.6584	0.6584	0.6497	0.6497	0.6651	0.6538	0.6610	0.6313	0.6560
IC models introduced by Lastra-Díaz and García-Serrano (2015a)													
CPHyponyms	0.6267	0.6466	0.6462	0.6329	0.6466	0.6466	0.6460	0.6460	0.6521	0.6491	0.6553	0.6455	0.6452
CPUniform	0.5617	0.6325	0.6005	0.5704	0.6325	0.6325	0.5973	0.6002	0.6482	0.6022	0.6455	0.6478	0.6395
CPLeaves	0.6289	0.6478	0.6476	0.6363	0.6478	0.6478	0.6476	0.6476	0.6534	0.6509	0.6553	0.6462	0.6485
CPCosine	0.6274	0.6479	0.6524	0.6388	0.6479	0.6479	0.6520	0.6524	0.6467	0.6548	0.6568	0.6458	0.6457
CPLogistic	0.6241	0.6460	0.6302	0.6284	0.6460	0.6460	0.6306	0.6302	0.6497	0.6325	0.6538	0.6423	0.6272
CPCorpus	0.6282	0.6364	0.6256	0.6202	0.6364	0.6364	0.6258	0.6256	0.6505	0.6284	0.6591	0.6458	0.6390
IC models introduced in this work													
CPRefHyponyms	0.6327	0.6495	0.6510	0.6424	0.6495	0.6495	0.6510	0.6510	0.6530	0.6528	0.6598	0.6468	0.6452
CPRefUniform	0.6384	0.6679	0.6236	0.6366	0.6679	0.6679	0.6236	0.6236	0.6486	0.6262	0.6665	0.6525	0.6395
CPRefLeaves	0.6344	0.6526	0.6551	0.6429	0.6526	0.6526	0.6551	0.6551	0.6538	0.6553	0.6609	0.6470	0.6485
CPRefCosine	0.6318	0.6459	0.6515	0.6416	0.6459	0.6459	0.6515	0.6515	0.6531	0.6536	0.6598	0.6463	0.6368
CPRefLogistic	0.6323	0.6538	0.6490	0.6355	0.6538	0.6538	0.6490	0.6490	0.6639	0.6515	0.6639	0.6578	0.6539
CPRefCorpus	0.6341	0.6435	0.6284	0.6297	0.6435	0.6435	0.6284	0.6284	0.6534	0.6328	0.6617	0.6504	0.6390
CPRefCosineLeaves	0.6314	0.6485	0.6537	0.6437	0.6485	0.6485	0.6537	0.6537	0.6529	0.6567	0.6598	0.6463	0.6419
CPRefLogisticLeaves	0.6312	0.6554	0.6509	0.6380	0.6554	0.6554	0.6509	0.6509	0.6650	0.6524	0.6638	0.6589	0.6558
CPRefLeavesSubsumersRatio	0.6424	0.6558	0.6540	0.6474	0.6558	0.6558	0.6540	0.6540	0.6539	0.6541	0.6590	0.6494	0.6475
Best column value	0.6481	0.6679	0.6612	0.6656	0.6679	0.6679	0.6612	0.6612	0.6656	0.6619	0.6678	0.6662	0.6670

Table 20: Spearman (ρ) correlation coefficients for all the IC models and measures in the Agirre201 dataset using WordNet 3.0. The bold values represent the best score within each column, whilst the bold underlined value is the best overall score in the dataset for any combination of IC-based similarity measure with any IC model. Last row shows the best values within each column.

P&S _{full} - r	Classic IC-based measures				Monotone trans. from classic measures				hybrid IC-based measures (shortest path length)				
	Resnik	Lin	J&C	P&S	FaITH	Meng12	Garla	cosJ&C	Li _{s9}	Zhou	Meng14	Gao	coswJ&C
Resnik (1995)	0.8736	0.8877	0.8749	0.8328	0.9003	0.8988	0.7761	0.8922	0.8857	0.8778	0.8286	0.8909	0.8826
Seco et al. (2004)	0.8799	0.8945	0.8781	0.8970	0.9042	0.9031	0.9021	0.8966	0.8839	0.8949	0.8374	0.8585	0.8819
Blanchard et al. (2008)	0.8788	0.8932	0.8736	0.8381	0.9020	0.9013	0.7813	0.8942	0.8890	0.8769	0.8375	0.8926	0.8811
Zhou et al. (2008a)	0.8357	0.8420	0.8372	0.8563	0.8905	0.8806	0.8979	0.8726	0.8897	0.8725	0.7488	0.8585	0.8593
Sánchez et al. (2011)	0.8740	0.8738	0.8762	0.8051	0.9025	0.8964	0.8470	0.8996	0.8807	0.8789	0.7943	0.8844	0.8850
Sánchez and Batet (2012)	0.8829	0.8948	0.8742	0.8398	0.9042	0.9035	0.7798	0.8918	0.8877	0.8771	0.8350	0.8920	0.8790
Meng et al. (2012)	0.8645	0.8863	0.8715	0.8897	0.9057	0.9025	0.9058	0.8917	0.8813	0.8805	0.8106	0.8585	0.8765
Yuan et al. (2013)	0.8655	0.8896	0.8641	0.8891	0.9082	0.9061	0.9010	0.8840	0.8774	0.8796	0.8085	0.8585	0.8713
Hadj Taieb et al. (2014a)	0.4915	0.7962	0.4261	0.4660	0.8167	0.8125	0.6749	0.4272	0.8724	0.4308	0.7632	0.8751	0.4272
Adhikari et al. (2015)	0.8653	0.8875	0.8735	0.8908	0.9056	0.9027	0.9057	0.8926	0.8804	0.8819	0.8163	0.8585	0.8774
IC models introduced by Lastra-Díaz and García-Serrano (2015a)													
CPHyponyms	0.8725	0.8903	0.8783	0.8325	0.9015	0.9002	0.7884	0.8963	0.8865	0.8811	0.8286	0.8899	0.8808
CPUniform	0.7077	0.8042	0.7204	0.6602	0.8644	0.8498	0.7575	0.7478	0.8759	0.7260	0.7300	0.8696	0.8169
CPLeaves	0.8715	0.8887	0.8748	0.8297	0.9008	0.8992	0.7865	0.8930	0.8864	0.8775	0.8277	0.8897	0.8777
CPCosine	0.8757	0.8915	0.8778	0.8318	0.9015	0.9005	0.7502	0.8943	0.8795	0.8797	0.8330	0.8807	0.8785
CPLogistic	0.8419	0.8928	0.8315	0.8061	0.9064	0.9056	0.7742	0.8632	0.8852	0.8363	0.8193	0.8880	0.8645
CPCorpus	0.8649	0.8840	0.8711	0.8203	0.8987	0.8964	0.8129	0.8892	0.8858	0.8744	0.8246	0.8906	0.8807
IC models introduced in this work													
CPRefHyponyms	0.8690	0.8870	0.8730	0.8312	0.8978	0.8966	0.7852	0.8918	0.8892	0.8763	0.8353	0.8912	0.8808
CPRefUniform	0.7577	0.8738	0.7511	0.7301	0.8979	0.8942	0.7564	0.7771	0.8763	0.7560	0.7915	0.8758	0.8169
CPRefLeaves	0.8678	0.8854	0.8694	0.8282	0.8971	0.8955	0.7835	0.8887	0.8892	0.8726	0.8343	0.8908	0.8777
CPRefCosine	0.8694	0.8893	0.8748	0.8311	0.8992	0.8982	0.7953	0.8903	0.8883	0.8787	0.8339	0.8912	0.8748
CPRefLogistic	0.8403	0.8709	0.8564	0.8142	0.8923	0.8885	0.7977	0.8797	0.8827	0.8599	0.8347	0.8751	0.8608
CPRefCorpus	0.8631	0.8824	0.8687	0.8216	0.8957	0.8935	0.8072	0.8862	0.8888	0.8722	0.8325	0.8919	0.8807
CPRefCosineLeaves	0.8684	0.8864	0.8689	0.8263	0.8978	0.8962	0.7898	0.8856	0.8882	0.8726	0.8323	0.8909	0.8706
CPRefLogisticLeaves	0.8407	0.8695	0.8558	0.8130	0.8920	0.8878	0.7950	0.8788	0.8821	0.8593	0.8328	0.8743	0.8603
CPRefLeavesSubsumersRatio	0.8740	0.8874	0.8677	0.8286	0.8989	0.8974	0.7696	0.8808	0.8890	0.8699	0.8349	0.8910	0.8721
Best column value	0.8829	0.8948	0.8783	0.8970	0.9082	0.9061	0.9058	0.8996	0.8897	0.8949	0.8375	0.8926	0.8850

Table 21: Pearson (r) correlation coefficients for all the IC models and measures in the $P\&S_{full}$ dataset using WordNet 3.0. The bold values represent the best score within each column, whilst the bold underlined value is the best overall score in the dataset for any combination of IC-based similarity measure with any IC model. Last row shows the best values within each column.

P&S _{full} - ρ	Classic IC-based measures				Monotone trans. from classic measures				hybrid IC-based measures (shortest path length)				
	Resnik	Lin	J&C	P&S	FaITH	Meng12	Garla	cosJ&C	Li _{s9}	Zhou	Meng14	Gao	coswJ&C
Resnik (1995)	0.7783	0.7660	0.7717	0.7422	0.7660	0.7660	0.7717	0.7717	0.7809	0.7737	0.8105	0.8089	0.7773
Seco et al. (2004)	0.7735	0.7911	0.7768	0.7919	0.7911	0.7911	0.7768	0.7768	0.7889	0.8144	0.8150	0.7986	0.7821
Blanchard et al. (2008)	0.7527	0.7788	0.7668	0.7629	0.7788	0.7788	0.7668	0.7668	0.7793	0.7737	0.8121	0.7899	0.7751
Zhou et al. (2008a)	0.7720	0.7867	0.7999	0.7938	0.7867	0.7867	0.7999	0.7999	0.7903	0.8212	0.8122	0.7986	0.8005
Sánchez et al. (2011)	0.7732	0.7936	0.8003	0.7573	0.7936	0.7936	0.8003	0.8003	0.7789	0.8031	0.8179	0.8046	0.8226
Sánchez and Batet (2012)	0.7727	0.7854	0.7652	0.7681	0.7854	0.7854	0.7652	0.7652	0.7806	0.7732	0.8116	0.7958	0.7740
Meng et al. (2012)	0.7543	0.7776	0.8127	0.8122	0.7776	0.7776	0.8127	0.8127	0.7842	0.8126	0.8072	0.7986	0.8215
Yuan et al. (2013)	0.7759	0.7905	0.7957	0.8199	0.7905	0.7905	0.7957	0.7957	0.7917	0.8138	0.8125	0.7986	0.7975
Hadj Taieb et al. (2014a)	0.6140	0.7463	0.6094	0.6107	0.7463	0.7463	0.6094	0.6094	0.7674	0.6153	0.7833	0.7798	0.6094
Adhikari et al. (2015)	0.7541	0.7770	0.8068	0.8045	0.7770	0.7770	0.8068	0.8068	0.7842	0.8122	0.8066	0.7986	0.8164
IC models introduced by Lastra-Díaz and García-Serrano (2015a)													
CPHyponyms	0.7461	0.7824	0.7910	0.7790	0.7824	0.7824	0.7910	0.7910	0.7744	0.7943	0.8103	0.7885	0.7819
CPUniform	0.6997	0.7852	0.7684	0.6905	0.7852	0.7852	0.7684	0.7684	0.7804	0.7696	0.7973	0.7833	0.7974
CPLeaves	0.7457	0.7808	0.7766	0.7689	0.7808	0.7808	0.7766	0.7766	0.7745	0.7833	0.8089	0.7884	0.7818
CPCosine	0.7689	0.7834	0.7710	0.7664	0.7834	0.7834	0.7710	0.7710	0.7788	0.7760	0.8118	0.8053	0.7791
CPLogistic	0.7783	0.7921	0.7880	0.7630	0.7921	0.7921	0.7880	0.7880	0.7811	0.7888	0.8075	0.7910	0.7695
CPCorpus	0.7691	0.7572	0.7615	0.7291	0.7572	0.7572	0.7615	0.7615	0.7809	0.7656	0.8116	0.8077	0.7628
IC models introduced in this work													
CPRefHyponyms	0.7545	0.7692	0.7640	0.7603	0.7692	0.7692	0.7640	0.7640	0.7789	0.7682	0.8061	0.7909	0.7819
CPRefUniform	0.7189	0.7752	0.7394	0.7124	0.7752	0.7752	0.7394	0.7394	0.7763	0.7434	0.8026	0.7782	0.7974
CPRefLeaves	0.7528	0.7704	0.7654	0.7600	0.7704	0.7704	0.7654	0.7654	0.7784	0.7738	0.8047	0.7774	0.7818
CPRefCosine	0.7520	0.7836	0.7763	0.7641	0.7836	0.7836	0.7763	0.7763	0.7819	0.7800	0.8134	0.7913	0.7909
CPRefLogistic	0.7288	0.7335	0.7513	0.7113	0.7335	0.7335	0.7513	0.7513	0.7425	0.7616	0.7626	0.7342	0.7742
CPRefCorpus	0.7539	0.7522	0.7460	0.7247	0.7522	0.7522	0.7460	0.7460	0.7804	0.7506	0.8027	0.7925	0.7628
CPRefCosineLeaves	0.7520	0.7710	0.7672	0.7585	0.7710	0.7710	0.7672	0.7672	0.7832	0.7728	0.8078	0.7931	0.7841
CPRefLogisticLeaves	0.7291	0.7327	0.7595	0.7140	0.7327	0.7327	0.7595	0.7595	0.7445	0.7635	0.7609	0.7334	0.7822
CPRefLeavesSubsumersRatio	0.7578	0.7582	0.7677	0.7507	0.7582	0.7582	0.7677	0.7677	0.7787	0.7734	0.7999	0.7872	0.7804
Best column value	0.7783	0.7936	0.8127	0.8199	0.7936	0.7936	0.8127	0.8127	0.7917	0.8212	0.8179	0.8089	0.8226

Table 22: Spearman (ρ) correlation coefficients for all the IC models and measures in the P&S_{full} dataset using WordNet 3.0. The bold values represent the best score within each column, whilst the bold underlined value is the best overall score in the dataset for any combination of IC-based similarity measure with any IC model. Last row shows the best values within each column.

SimLex665 - r	Classic IC-based measures				Monotone trans. from classic measures				hybrid IC-based measures (shortest path length)				
	Resnik	Lin	J&C	P&S	FaITH	Meng12	Garla	cosJ&C	Li _{s9}	Zhou	Meng14	Gao	coswJ&C
Resnik (1995)	0.5355	0.5935	0.5692	0.5489	0.5948	0.5955	0.4644	0.5781	0.6110	0.5753	0.5659	0.6152	0.5775
Seco et al. (2004)	0.5339	0.6010	0.5918	0.5975	0.6046	0.6048	0.6035	0.6013	0.5763	0.6237	0.5712	0.5602	0.6027
Blanchard et al. (2008)	0.5320	0.5958	0.5828	0.5517	0.5986	0.5990	0.4653	0.5935	0.6133	0.5882	0.5696	0.6154	0.5960
Zhou et al. (2008a)	0.5110	0.5778	0.5825	0.5803	0.5985	0.5945	0.6092	0.5973	0.5912	0.6162	0.5300	0.5602	0.6051
Sánchez et al. (2011)	0.5364	0.5791	0.5838	0.5318	0.5904	0.5883	0.5117	0.5941	0.6064	0.5887	0.5448	0.6112	0.5918
Sánchez and Batet (2012)	0.5354	0.5964	0.5861	0.5532	0.5991	0.5995	0.4637	0.5945	0.6124	0.5910	0.5683	0.6153	0.5976
Meng et al. (2012)	0.4972	0.5841	0.5939	0.5863	0.5887	0.5884	0.6064	0.6010	0.5661	0.6149	0.5526	0.5602	0.6056
Yuan et al. (2013)	0.5120	0.6004	0.6027	0.5981	0.6063	0.6062	0.6166	0.6106	0.5722	0.6203	0.5579	0.5602	0.6157
Hadj Taieb et al. (2014a)	0.2033	0.4849	0.2154	0.3237	0.4921	0.4896	0.3572	0.2177	0.5359	0.2192	0.5078	0.5448	0.2177
Adhikari et al. (2015)	0.4983	0.5848	0.5945	0.5871	0.5893	0.5890	0.6069	0.6016	0.5648	0.6155	0.5541	0.5602	0.6060
IC models introduced by Lastra-Díaz and García-Serrano (2015a)													
CPHyponyms	0.5223	0.5910	0.5811	0.5468	0.5932	0.5940	0.4573	0.5896	0.6118	0.5867	0.5660	0.6144	0.6023
CPUniform	0.3632	0.5245	0.5366	0.4693	0.5186	0.5260	0.3908	0.5416	0.6044	0.5408	0.5091	0.5992	0.6010
CPLeaves	0.5204	0.5904	0.5808	0.5457	0.5927	0.5934	0.4565	0.5897	0.6118	0.5864	0.5652	0.6143	0.6033
CPCosine	0.5274	0.5934	0.5841	0.5490	0.5959	0.5964	0.4169	0.5920	0.6063	0.5872	0.5688	0.6109	0.5984
CPLogistic	0.5068	0.5926	0.5679	0.5441	0.5966	0.5972	0.3876	0.5778	0.6098	0.5726	0.5589	0.6141	0.5976
CPCorpus	0.5271	0.5856	0.5564	0.5404	0.5844	0.5863	0.0593	0.5642	0.6112	0.5639	0.5630	0.6152	0.5759
IC models introduced in this work													
CPRefHyponyms	0.5204	0.5935	0.5864	0.5525	0.5962	0.5966	0.4689	0.5950	0.6133	0.5918	0.5704	0.6141	0.6023
CPRefUniform	0.4373	0.5704	0.5563	0.5180	0.5719	0.5728	0.4269	0.5612	0.6046	0.5599	0.5507	0.6050	0.6010
CPRefLeaves	0.5187	0.5931	0.5869	0.5519	0.5962	0.5965	0.4693	0.5958	0.6132	0.5921	0.5697	0.6138	0.6033
CPRefCosine	0.5207	0.5966	0.5875	0.5557	0.5978	0.5987	0.4718	0.5943	0.6128	0.5931	0.5723	0.6141	0.5941
CPRefLogistic	0.4940	0.5639	0.5744	0.5349	0.5781	0.5758	0.4519	0.5824	0.6039	0.5787	0.5522	0.6003	0.5895
CPRefCorpus	0.5268	0.5886	0.5611	0.5453	0.5883	0.5895	0.4682	0.5693	0.6132	0.5683	0.5677	0.6152	0.5759
CPRefCosineLeaves	0.5188	0.5963	0.5889	0.5548	0.5984	0.5990	0.4735	0.5964	0.6129	0.5943	0.5712	0.6139	0.5973
CPRefLogisticLeaves	0.4914	0.5636	0.5744	0.5343	0.5785	0.5760	0.4526	0.5828	0.6033	0.5786	0.5509	0.5996	0.5903
CPRefLeavesSubsumersRatio	0.5124	0.5908	0.5837	0.5455	0.5922	0.5931	0.4541	0.5912	0.6131	0.5878	0.5670	0.6139	0.6013
Best column value	0.5364	0.6010	0.6027	0.5981	0.6063	0.6062	0.6166	0.6106	0.6133	0.6237	0.5723	0.6154	0.6157

Table 23: Pearson (r) correlation coefficients for all the IC models and measures in the SimLex665 dataset using WordNet 3.0. The bold values represent the best score within each column, whilst the bold underlined value is the best overall score in the dataset for any combination of IC-based similarity measure with any IC model. Last row shows the best values within each column.

SimLex665 - ρ	Classic IC-based measures				Monotone trans. from classic measures				hybrid IC-based measures (shortest path length)				
	Resnik	Lin	J&C	P&S	FaITH	Meng12	Garla	cosJ&C	Li _{s9}	Zhou	Meng14	Gao	coswJ&C
Resnik (1995)	0.5210	0.5810	0.5700	0.5450	0.5810	0.5810	0.5700	0.5700	0.5920	0.5741	0.6057	0.5994	0.5683
Seco et al. (2004)	0.5210	0.5888	0.5901	0.5862	0.5888	0.5888	0.5901	0.5901	0.5809	0.6101	0.6105	0.5918	0.5916
Blanchard et al. (2008)	0.5190	0.5844	0.5820	0.5466	0.5844	0.5844	0.5820	0.5820	0.5892	0.5853	0.6071	0.5955	0.5848
Zhou et al. (2008a)	0.5026	0.5841	0.5945	0.5767	0.5841	0.5841	0.5945	0.5945	0.5804	0.6089	0.6025	0.5918	0.6001
Sánchez et al. (2011)	0.5248	0.5803	0.5906	0.5204	0.5803	0.5803	0.5906	0.5906	0.5888	0.5933	0.6015	0.5953	0.5873
Sánchez and Batet (2012)	0.5232	0.5850	0.5835	0.5476	0.5850	0.5850	0.5835	0.5835	0.5896	0.5866	0.6068	0.5965	0.5868
Meng et al. (2012)	0.4896	0.5767	0.5957	0.5714	0.5767	0.5767	0.5957	0.5957	0.5672	0.6050	0.5964	0.5918	0.5980
Yuan et al. (2013)	0.5030	0.5903	0.6027	0.5869	0.5903	0.5903	0.6027	0.6027	0.5727	0.6114	0.6051	0.5918	0.6076
Hadj Taieb et al. (2014a)	0.3004	0.4833	0.3256	0.4243	0.4833	0.4833	0.3256	0.3256	0.5275	0.3307	0.5432	0.5423	0.3256
Adhikari et al. (2015)	0.4899	0.5766	0.5960	0.5718	0.5766	0.5766	0.5960	0.5960	0.5673	0.6050	0.5968	0.5918	0.5983
IC models introduced by Lastra-Díaz and García-Serrano (2015a)													
CPHyponyms	0.5067	0.5799	0.5806	0.5391	0.5798	0.5799	0.5723	0.5806	0.5873	0.5847	0.6046	0.5920	0.5922
CPUniform	0.3649	0.5223	0.5506	0.4620	0.5220	0.5223	0.5672	0.5506	0.5825	0.5530	0.5572	0.5846	0.6028
CPLeaves	0.5043	0.5797	0.5799	0.5392	0.5796	0.5797	0.5716	0.5799	0.5867	0.5839	0.6044	0.5924	0.5929
CPCosine	0.5150	0.5821	0.5828	0.5413	0.5821	0.5821	0.5745	0.5827	0.5898	0.5849	0.6070	0.5924	0.5884
CPLogistic	0.4925	0.5791	0.5738	0.5358	0.5790	0.5791	0.5738	0.5738	0.5885	0.5767	0.6032	0.5936	0.5904
CPCorpus	0.5085	0.5735	0.5578	0.5335	0.5735	0.5735	0.5460	0.5578	0.5912	0.5639	0.6024	0.5983	0.5662
IC models introduced in this work													
CPRefHyponyms	0.5077	0.5822	0.5856	0.5470	0.5822	0.5822	0.5856	0.5856	0.5896	0.5894	0.6076	0.5946	0.5922
CPRefUniform	0.4438	0.5599	0.5639	0.5097	0.5599	0.5599	0.5639	0.5639	0.5861	0.5662	0.5931	0.5914	0.6028
CPRefLeaves	0.5073	0.5817	0.5864	0.5469	0.5817	0.5817	0.5864	0.5864	0.5891	0.5901	0.6074	0.5943	0.5929
CPRefCosine	0.5081	0.5844	0.5864	0.5509	0.5844	0.5844	0.5864	0.5864	0.5879	0.5906	0.6099	0.5920	0.5863
CPRefLogistic	0.4876	0.5646	0.5763	0.5231	0.5646	0.5646	0.5763	0.5763	0.5878	0.5797	0.6007	0.5902	0.5803
CPRefCorpus	0.5084	0.5762	0.5612	0.5407	0.5762	0.5762	0.5612	0.5612	0.5929	0.5667	0.6045	0.5998	0.5662
CPRefCosineLeaves	0.5056	0.5840	0.5881	0.5498	0.5840	0.5840	0.5881	0.5881	0.5894	0.5915	0.6089	0.5927	0.5897
CPRefLogisticLeaves	0.4869	0.5657	0.5777	0.5235	0.5657	0.5657	0.5777	0.5777	0.5868	0.5808	0.6004	0.5897	0.5817
CPRefLeavesSubsumersRatio	0.5112	0.5785	0.5859	0.5466	0.5785	0.5785	0.5859	0.5859	0.5900	0.5884	0.6030	0.5952	0.5946
Best column value	0.5248	0.5903	0.6027	0.5869	0.5903	0.5903	0.6027	0.6027	0.5929	0.6114	0.6105	0.5998	0.6076

Table 24: Spearman (ρ) correlation coefficients for all the IC models and measures in the SimLex665 dataset using WordNet 3.0. The bold values represent the best score within each column, whilst the bold underlined value is the best overall score in the dataset for any combination of IC-based similarity measure with any IC model. Last row shows the best values within each column.

IC models	Resnik (1995)	Seco et al. (2004)	Blanchard et al. (2008)	Zhou et al. (2008a)	Sánchez et al. (2011)	Sánchez and Batet (2012)	Meng et al. (2012)	Yuan et al. (2013)	Hadj Taitab et al. (2014a)	Adhikari et al. (2015)	CPHyponyms	CPUniform	CPLeaves	CPCosine	CPLogistic	CPCorpus	CPRetHyponyms	CPRetUniform	CPRetLeaves	CPRetCosine	CPRetLogistic	CPRetCorpus	CPRetCosineLeaves	CPRetLogisticLeaves	CPRetLeavesSubsumersRatio
Resnik	.008	.496	.019	.122	.368	.061	.020	.000	.084	.420	.000	.038	.409	.000	.000	.000	.422	.000	.414	.133	.000	.001	.332	.000	.011
Seco et al	.008	—	.020	.149	.001	.036	.005	.000	.010	.000	.000	.000	.000	.000	.000	.000	.002	.000	.001	.003	.000	.001	.001	.000	.001
Blanchard	.496	.000	.001	.172	.111	.024	.001	.000	.031	.387	.000	.001	.328	.000	.000	.000	.375	.000	.289	.003	.000	.002	.211	.000	.006
Zhou et al	.019	.020	.001	.306	.003	.329	.173	.000	.053	.000	.000	.000	.000	.000	.000	.000	.004	.000	.002	.007	.000	.001	.001	.000	.001
Sánchez11	.122	.149	.172	.306	—	.362	.256	.000	.444	.124	.000	.029	.135	.002	.001	.000	.156	.000	.144	.297	.000	.006	.140	.000	.027
Sánchez12	.368	.001	.111	.003	.219	—	.041	.002	.058	.278	.000	.001	.170	.000	.000	.000	.212	.000	.134	.041	.000	.001	.115	.000	.004
Meng et al	.061	.036	.024	.329	.362	.041	—	.159	.010	.002	.000	.001	.010	.002	.001	.000	.028	.000	.026	.079	.000	.003	.017	.000	.007
Yuan et al	.020	.005	.001	.173	.256	.002	.159	—	.056	.001	.000	.000	.001	.000	.000	.000	.005	.000	.004	.010	.000	.001	.002	.000	.001
H.Taitab	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.001	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.001	.000
Adhikari	.084	.010	.031	.053	.444	.058	.010	.056	—	.003	.000	.000	.013	.002	.001	.000	.037	.000	.033	.118	.000	.003	.021	.000	.008
CPHypo	.420	.000	.387	.000	.124	.278	.002	.001	.000	.003	—	.020	.465	.006	.003	.468	.001	.000	.479	.096	.000	.019	.467	.000	.084
CPUnif	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	—	.000	.000	.001	.000	.000	.021	.000	.000	.013	.001	.000	.010	.001
CPLeaves	.038	.000	.001	.000	.029	.001	.001	.000	.000	.020	.000	.000	.001	.013	.004	.001	.013	.001	.005	.000	.000	.025	.010	.000	.218
CPCosine	.409	.000	.328	.000	.135	.170	.010	.001	.013	.465	.000	.001	—	.001	.001	.001	.479	.000	.499	.025	.000	.004	.407	.000	.027
CPLogistic	.000	.000	.000	.002	.000	.002	.000	.000	.002	.006	.000	.001	.001	—	.001	.092	.000	.002	.000	.000	.001	.418	.000	.006	.013
CPCorpus	.000	.000	.000	.001	.000	.001	.000	.000	.001	.003	.001	.004	.001	.092	.000	—	.000	.003	.000	.000	.002	.029	.000	.015	.003
CPRetHypo	.422	.002	.375	.004	.156	.212	.028	.005	.000	.037	.468	.000	.13	.479	.000	.000	—	.000	.459	.025	.000	.000	.295	.000	.001
CPRetUnif	.000	.000	.000	.000	.000	.000	.000	.000	.000	.001	.001	.021	.001	.000	.002	.003	.000	.000	.000	.000	.110	.001	.000	.084	.000
CPRetLea	.414	.001	.289	.002	.144	.134	.026	.004	.000	.033	.479	.000	.005	.499	.000	.000	.459	.000	—	.000	.000	.001	.371	.000	.002
CPRetCos	.133	.003	.003	.007	.297	.041	.079	.010	.118	.096	.000	.013	.025	.000	.000	.000	.025	.000	.007	—	.000	.000	.001	.000	.001
CPRetLog	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.001	.002	.000	.110	.000	.000	—	.002	.000	.009	.000
CPRetCorpus	.001	.001	.002	.001	.006	.001	.003	.001	.003	.019	.001	.025	.004	.418	.029	.000	.029	.000	.001	.000	.002	—	.001	.011	.006
CPRetCosLea	.332	.001	.211	.001	.140	.115	.017	.002	.000	.021	.467	.000	.407	.000	.006	.015	.295	.000	.371	.001	.000	.001	—	.000	.007
CPRetLogLea	.000	.000	.000	.000	.000	.000	.000	.000	.001	.000	.000	.010	.000	.000	.006	.006	.084	.000	.000	.000	.009	.011	.000	—	.001
CPRetLeSuRat	.011	.001	.006	.001	.027	.004	.007	.001	.008	.084	.001	.218	.027	.013	.003	.001	.001	.000	.002	.001	.000	.006	.007	.000	—

Table 25: Pairwise p-values of the one-sided t-Student distribution for the paired difference between the average Spearman (ρ) correlation values of each pair of IC models on all the IC-based similarity measures. Each pairwise p-value is computed using the vector of average Spearman correlation values of each IC model with each IC-based similarity measure (rows in table 11) as paired random sample set. For a level of significance of 5%, each p-value ≤ 0.05 denotes a statistically significant higher or lower performance between these two IC models.

Simi. measures	Lastra & Garcia (2015), coswJ&C	Zhou et al (2008)	Lastra & Garcia (2015), cosJ&C	P&S (2008)	Taieb et al (2014)	Gao et al (2015)	J&C (1997)	Meng & Gu (2012)	P&S FaITH (2010)	Lin (1998)	Li et al (2003) Strat. 3	Li _{s9} (2003)	Li et al (2003) Strat. 4	Sánchez et al (2012)	Meng et al (2014)	Al-Mubaid & Nguyen (2009)	Resnik (1995)	Pedersen et al (2007)	Leacock & Chodorow (1998)	Garla	Rada et al (1989)	Wu & Palmer (1994)	
L&G, coswJ&C	.026	.248	.026	.045	.067	.084	.026	.020	.020	.020	.029	.034	.042	.024	.178	.038	.003	.029	.029	.029	.026	.029	.006
Zhou et al	.248	—	.034	.184	.023	.028	.034	.002	.002	.002	.002	.012	.015	.006	.149	.019	.020	.002	.002	.002	.034	.002	.002
L&G, cosJ&C	.026	.034	—	.346	.207	.338	1.00	.023	.023	.023	.129	.083	.088	.042	.372	.188	.020	.129	.129	1.00	1.00	.129	.005
P&S	.045	.184	.346	—	.239	.340	.346	.114	.114	.114	.185	.124	.123	.079	.463	.226	.007	.185	.185	.346	.346	.185	.010
H.Taieb	.067	.023	.207	.239	—	.294	.207	.499	.499	.499	.363	.016	.036	.001	.022	.413	.096	.363	.363	.207	.363	.363	.000
Gao et al	.084	.028	.338	.340	.294	—	.338	.345	.345	.345	.077	.087	.015	.020	.029	.163	.095	.077	.077	.338	.077	.338	.004
J&C	.026	.034	1.00	.346	.207	.338	—	.023	.023	.023	.129	.083	.088	.042	.372	.188	.020	.129	.129	1.00	1.00	.129	.005
Meng & Gu	.020	.002	.023	.114	.499	.345	.023	—	1.00	1.00	.346	.166	.132	.064	.104	.430	.055	.346	.346	.023	.346	.346	.005
P&S, FaITH	.020	.002	.023	.114	.499	.345	.023	1.00	—	1.00	.346	.166	.132	.064	.104	.430	.055	.346	.346	.023	.346	.346	.005
Lin	.020	.002	.023	.114	.499	.345	.023	1.00	1.00	—	.346	.166	.132	.064	.104	.430	.055	.346	.346	.023	.346	.346	.005
Li _{s3} et al	.029	.002	.129	.185	.363	.077	.129	.346	.346	.346	—	.313	.132	.130	.034	.411	.124	1.00	1.00	.129	.129	1.00	.013
Li _{s4} et al	.034	.012	.083	.124	.016	.087	.083	.166	.166	.166	.313	—	.168	.019	.006	.166	.145	.313	.313	.083	.083	.313	.000
Sánchez et al	.042	.015	.088	.123	.036	.015	.088	.132	.132	.132	.132	.168	—	.443	.001	.069	.281	.132	.132	.088	.088	.132	.007
Meng et al	.024	.006	.042	.079	.001	.020	.042	.064	.064	.064	.130	.019	.443	—	.001	.044	.226	.130	.130	.042	.042	.130	.001
Al-Mubaid	.178	.149	.372	.463	.022	.029	.372	.104	.104	.104	.034	.006	.001	.001	—	.018	.056	.034	.034	.372	.034	.034	.001
Resnik	.038	.019	.188	.226	.413	.163	.188	.430	.430	.430	.411	.166	.069	.044	.018	—	.099	.411	.411	.188	.411	.188	.004
Pedersen	.003	.020	.020	.007	.096	.095	.020	.055	.055	.055	.124	.145	.281	.226	.056	.099	—	.124	.124	.020	.020	.124	.203
L & Ch.	.029	.002	.129	.185	.363	.077	.129	.346	.346	.346	1.00	.313	.132	.130	.034	.411	.124	—	1.00	.129	.129	1.00	.013
Garla & Brandt	.029	.002	.129	.185	.363	.077	.129	.346	.346	.346	1.00	.313	.132	.130	.034	.411	.124	1.00	—	.129	.129	1.00	.013
Rada et al	.026	.034	1.00	.346	.207	.338	1.00	.023	.023	.023	.129	.083	.088	.042	.372	.188	.020	.129	.129	—	—	.129	.005
W & P	.029	.002	.129	.185	.363	.077	.129	.346	.346	.346	1.00	.313	.132	.130	.034	.411	.124	1.00	1.00	.129	.129	—	.013
	.006	.002	.005	.010	.000	.004	.005	.005	.005	.005	.013	.000	.007	.001	.001	.004	.203	.013	.013	.005	.005	.013	—

Table 26: Pairwise p-values of the one-sided t-Student distribution for the paired difference between the Spearman (ρ) correlation values of each pair of similarity measures on all datasets. Each pairwise p-value is computed using the vector of average Spearman correlation values of each similarity measure on all datasets (rows in table 12) as paired random sample set. For a level of significance of 5%, each p-value ≤ 0.05 denotes a statistically significant higher or lower performance between these two similarity measures.

Chapter 10

Information Systems article

This page intentionally left blank.



HESML: A scalable ontology-based semantic similarity measures library with a set of reproducible experiments and a replication dataset



Juan J. Lastra-Díaz^{a,*}, Ana García-Serrano^a, Montserrat Batet^{b,1}, Miriam Fernández^{c,1}, Fernando Chirigati^{d,1}

^a NLP & IR Research Group, E.T.S.I. Informática, Universidad Nacional de Educación a Distancia (UNED), C/Juan del Rosal 16, 28040 Madrid, (Spain)

^b Internet Interdisciplinary Institute (IN3), Universitat Oberta de Catalunya, Av. Carl Friedrich Gauss, 5. 08860 Castelldefels, Spain

^c Knowledge Media Institute, The Open University, Walton Hall, Milton Keynes, MK7 6AA, United Kingdom

^d Department of Computer Science and Engineering, New York University, New York City, NY, United States

ARTICLE INFO

Article history:

Received 26 July 2016

Revised 6 February 2017

Accepted 10 February 2017

Available online 21 February 2017

Keywords:

HESML

PosetHERep

Semantic measures library

Ontology-based semantic similarity measures

Intrinsic and corpus-based Information Content models

Reproducible experiments on word similarity

WNSimRep v1 dataset

ReproZip

WordNet-based semantic similarity measures

ABSTRACT

This work is a detailed companion reproducibility paper of the methods and experiments proposed by Lastra-Díaz and García-Serrano in (2015, 2016) [56–58], which introduces the following contributions: (1) a new and efficient representation model for taxonomies, called *PosetHERep*, which is an adaptation of the half-edge data structure commonly used to represent discrete manifolds and planar graphs; (2) a new Java software library called the *Half-Edge Semantic Measures Library (HESML)* based on *PosetHERep*, which implements most ontology-based semantic similarity measures and Information Content (IC) models reported in the literature; (3) a set of reproducible experiments on word similarity based on *HESML* and *ReproZip* with the aim of exactly reproducing the experimental surveys in the three aforementioned works; (4) a replication framework and dataset, called *WNSimRep v1*, whose aim is to assist the exact replication of most methods reported in the literature; and finally, (5) a set of scalability and performance benchmarks for semantic measures libraries. *PosetHERep* and *HESML* are motivated by several drawbacks in the current semantic measures libraries, especially the performance and scalability, as well as the evaluation of new methods and the replication of most previous methods. The reproducible experiments introduced herein are encouraged by the lack of a set of large, self-contained and easily reproducible experiments with the aim of replicating and confirming previously reported results. Likewise, the *WNSimRep v1* dataset is motivated by the discovery of several contradictory results and difficulties in reproducing previously reported methods and experiments. *PosetHERep* proposes a memory-efficient representation for taxonomies which linearly scales with the size of the taxonomy and provides an efficient implementation of most taxonomy-based algorithms used by the semantic measures and IC models, whilst *HESML* provides an open framework to aid research into the area by providing a simpler and more efficient software architecture than the current software libraries. Finally, we prove the outperformance of *HESML* on the state-of-the-art libraries, as well as the possibility of significantly improving their performance and scalability without caching using *PosetHERep*.

© 2017 The Authors. Published by Elsevier Ltd.

This is an open access article under the CC BY-NC-ND license. (<http://creativecommons.org/licenses/by-nc-nd/4.0/>)

1. Introduction

Human similarity judgments between concepts underlie most of cognitive capabilities, such as categorization, memory, decision-making and reasoning. Thus, the proposal for concept similarity models to estimate the degree of similarity between word and concept pairs has been a very active line of research in the fields of cognitive sciences [106,124], artificial intelligence and Information Retrieval (IR) [107]. The semantic similarity measures esti-

* Corresponding author.

E-mail addresses: jlastra@invi.uned.es (J.J. Lastra-Díaz), agarcia@lsi.uned.es

(A. García-Serrano), mbatetsa@uoc.edu (M. Batet), m.fernandez@open.ac.uk

(M. Fernández), fchirigati@nyu.edu (F. Chirigati).

¹ Reviewers

mates the degree of similarity between concepts by considering only ‘is-a’ relationships, whilst the semantic relatedness measures also consider any type of co-occurrence relationship. For instance, a *wheel* is closely related to a *car* because the wheels are part of any car; however, a *wheel* neither is a car nor derives from another common close concept as *vehicle*, thus their degree of similarity is low. Whilst hand-coded taxonomies, such as WordNet and other sources of knowledge, can be efficiently and reliably used to retrieve the ‘is-a’ relationships between concepts and words, the co-occurrence relationships required by the semantic relatedness measures need to be retrieved from a large corpus. For this reason [57, §1.1], ontology-based semantic similarity measures exclusively based on ‘is-a’ relationships are currently the best and most reliable strategy to estimate the degree of similarity between words and concepts [58], whilst the corpus-based similarity measures are the best strategy for estimating their degree of relatedness [8].

An ontology-based semantic similarity measure is a binary concept-valued function $sim : C \times C \rightarrow \mathbb{R}$ defined on a single-root taxonomy of concepts (C, \leq_C) , which returns an estimation of the degree of similarity between concepts as perceived by a human being. The ontology-based similarity measures have become both a very active research topic, and a key component in many applications. For instance, in the fields of Natural Language Processing (NLP) and IR, ontology-based semantic similarity measures have been used in Word Sense Disambiguation (WSD) methods [92], text similarity measures [86], spelling error detection [20], sentence similarity models [44,66,91], paraphrase detection [36], unified sense disambiguation methods for different types of structured sources of knowledge [73], document clustering [31], ontology alignment [30], document [74] and query anonymization [11], clustering of nominal information [9,10], chemical entity identification [40], interoperability among agent-based systems [34], and ontology-based Information Retrieval (IR) models [55,62] to solve the lack of an intrinsic semantic distance in vector ontology-based IR models [23]. In the field of bioengineering, ontology-based similarity measures have been proposed for synonym recognition [24] and biomedical text mining [14,98,112]. However, since the pioneering work of Lord et al. [72], the proposal of similarity measures for genomics and proteomics based on the Gene Ontology (GO) [5] have attracted a lot of attention, as detailed in a recent survey on the topic [76]. Many GO-based semantic similarity measures have been proposed for protein functional similarity [28,29,101,132], giving rise to applications in protein classification and protein-protein interactions [41,129], gene prioritization [117] and many others reported in [76, p.2].

In [57], Lastra-Díaz and García-Serrano introduce a new family of similarity measures based on an Information Content (IC) model, whose pioneering work is introduced by Resnik [108]. Their new family of semantic similarity measures is based on two unexplored notions: a non-linear normalization of the classic Jiang-Conrath distance [52], and a generalization of this latter distance on non tree-like taxonomies defined as the length of the shortest path within an IC-weighted taxonomy. One of the similarity measures introduced in [57], called *coswJ&Csim*, obtains the best results on the RG65 dataset. In another subsequent work [56], the same aforementioned authors introduce a new family of intrinsic and corpus-based IC models and a new algebraic framework for their derivation, which is based on the estimation of the conditional probabilities between child and parent concepts within a taxonomy. This latter family of IC models is refined in another subsequent paper [58], which also sets out the new state of the art and confirms the outperformance of the *coswJ&Csim* similarity measure in a statistically significant manner among the family of ontology-based semantic similarity measures based on WordNet.

Given a taxonomy of concepts defined by the triplet $C = ((C, \leq_C), \Gamma)$, where $\Gamma \in C$ is the supreme element called the root, an Information Content model is a function $IC : C \rightarrow \mathbb{R}^+ \cup \{0\}$,

which represents an estimation of the information content for every concept, defined by $IC(c_i) = -\log_2(p(c_i))$, $p(c_i)$ being the occurrence probability of each concept $c_i \in C$. Each IC model must satisfy two further properties: (1) nullity in the root, such that $IC(\Gamma) = 0$, and (2) growing monotonicity from the root to the leaf concepts, such that $\forall c_i \leq c_j \Rightarrow IC(c_i) \geq IC(c_j)$. Once the IC-based measure is chosen, the IC model is mainly responsible for the definition of the notion of similarity and distance between concepts.

The main aim of this work is to introduce the *PosetHERep* representation model and make the *Half-Edge Semantic Measures Library* (HESML) publicly available for the first time, together with a set of reproducible experiments whose aims are the exact replication of the three aforementioned experimental surveys [56–58], as well as the proposal for a self-contained experimental platform which can be easily used for extensive experimentation, even with no software coding. In addition, this work also introduces a new replication framework and the *WNSimRep v1* dataset for the first time provided as supplementary material in [63], whose aim is to provide a gold standard to assist in the exact replication of ontology-based similarity measures and IC models. Finally, we have carried-out a series of experiments in order to evaluate the scalability and performance of HESML as regards the Semantic Measures Library (SML) [48] and WNetSS [15], which sets out the current state of the art. This work is part of a novel initiative on computational reproducibility recently introduced by Chirigati et al. [26], whose pioneering work is introduced by Wolke et al. [127] with the aim of leading to the exact replication of several dynamic resource allocation strategies in cloud data centers evaluated in a companion paper [128].

1.1. Main motivation and hypothesis

The two main motivations of this work are three drawbacks in the current semantic measures libraries, detailed below, and the lack of a set of self-contained and easily reproducible experiments into ontology-based semantic similarity measures and IC models based on WordNet. Another significant motivation, also related to the reproducibility, is the lack of a gold standard to assist in the exact replication of ontology-based similarity measures and IC models.

1.1.1. On the current semantic measures libraries

Our first motivation is the discovery of several scalability and performance drawbacks in the current state-of-the-art semantic measures libraries. We argue that these aforementioned drawbacks are derived from the use of naive graph representation models which do not capture the intrinsic structure of the taxonomies being represented. As a consequence of this latter fact, all topological algorithms based on naive representation models demand a high computational cost which degrades their performance. In turn, in order to solve the performance problem of their graph-based algorithms, the current semantic measures libraries adopt a caching strategy, storing the ancestors and descendant sets of all vertexes within the taxonomy, among other topological queries in memory. This latter caching strategy significantly increases the memory usage and leads to a scalability problem as regards the size of the taxonomy, in addition to impacting the performance because of the further memory allocation and dynamic resizing of the caching data structures, or the interrogation of external relational databases.

Our main hypothesis is that a new representation model for taxonomies which properly encodes their intrinsic structure, together with a new software library based on it, should bridge the aforementioned gap of scalability and performance of the current semantic measures libraries. Thus, *our main research questions* are as follows: (Q1) is a new intrinsic representation model for taxonomies able to improve significantly the performance and scala-

bility of the state-of-the-art semantic measures libraries?, and (Q2) is it possible to significantly improve the performance and scalability of the state-of-the-art semantic measures libraries without using any caching strategy?

The current state-of-the-art libraries are based on caching for most topological queries and the *delocalization of attributes* from their base objects (vertexes and edges). For instance, SML represents the ontologies by graphs, in which each vertex and oriented edge is defined by a URI key in a Java hash set. Thus, any further information associated to each vertex or edge needs to be stored in any independent external data structure, an approach that we call *delocalized attributes*. In addition, SML uses hash sets to store all pre-computed information and topological queries associated to each vertex as follows: its incoming and outgoing edge sets, its ascendant and descendant sets, its minimum and maximum depths, its subsumed leaves and its IC values, among others. Following the same *delocalized approach*, the edge weights in SML are also stored in Java hash sets indexed by edge URIs. All the aforementioned taxonomical features are computed during the pre-processing step, or the first time that they are requested, being stored in their corresponding caching structures defined as hash sets or tables. All topological queries, as well as the shortest path algorithm implemented by SML, are based on the traversal of the SML graph model, as well as the cache information of the vertexes and their delocalized attributes. The cached taxonomical features are represented in a distributed collection of hash maps and sets indexed by edge and vertex URI keys. In short, the entire topological model of the SML is based on caching, hash maps and delocalized attributes from their base objects. One of the first consequences of caching the vertex sets, as the ancestor or descendant sets, is that it implies a non-linear increase in the use of memory. On the other hand, the delocalized approach adds a performance penalty because of the need to interrogate different hash maps in order to retrieve multiple attributes from the same underlying object, in addition to an increase in the memory required derived from the internal searching and storing structures required by the underlying hash maps. Finally, all graph traversal algorithms, especially the shortest path computation, suffer a significant decrease in performance derived from the lack of an efficient representation of the adjacency model. The SML algorithms need to interrogate the hash maps continuously by storing the incoming and outgoing edge sets of each vertex in order to retrieve the adjacency information and traverse the graph. Thus, the traversing method is especially time consuming in complex algorithms as the shortest path computation. Another significant example of caching is the approach adopted by the WNetSS semantic measures library introduced recently by Aouicha et al. [15]. Unlike SML, which computes the topological features on-the-fly by storing them in an in-memory cache, WNetSS carries-out a time-consuming off-line pre-processing of all WordNet-based topological information which is stored in a MySQL server. This latter caching strategy based on MySQL could be appropriate for supporting a large Web-based experimental platform, such as the SISR system proposed in [15]. However, it severely impacts the performance, scalability and extensibility of WNetSS.

A second motivation is related to several software architecture issues that lead to practical difficulties for the functional extension of current software libraries. For instance, WordNet::Similarity [99] and WS4J [121] were designed before the emergence of the intrinsic IC models described in Section 2.1, thus, these libraries maintain in-memory tables with the concept frequency counts which are interrogated in order to compute the IC values required in a similarity evaluation step; however, their data structures do not provide any proper abstraction layer or software architecture to integrate new intrinsic IC models easily. On the other hand, SML separates the in-memory storage of the IC values and edge weights from the edge and nodes within the base taxonomy by defining

two Java abstract interfaces to integrate new weighting schemes and IC models as external data providers which are interrogated on-the-fly. This latter software design decision looks fine from an abstract point of view; however, it hinders the implementation of weighted IC-based measures like the weighted J&C and coswJ&C similarity measures introduced by Lastra-Díaz and García-Serrano [57], because the edge weights depend on the IC values of the nodes.

A third motivation is the lack of software implementations for the most recent ontology-based similarity measures and intrinsic IC models developed during the last decade. This latter fact prevents the publication of exhaustive experimental surveys comparing the new proposed methods with most recent methods reported in the literature, because of the effort and difficulty in replicating previous methods and experiments.

1.1.2. On the reproducibility in the area

A fourth motivation of this work is the lack of a set of self-contained and easily reproducible experiments that allow the research community to be able to replicate methods and results reported in the literature exactly, even without the need for software coding. The lack of reproducible experiments, together with the aforementioned lack of software libraries covering the most recent methods, and the difficulties in replicating methods and experiments exactly have contributed, with few exceptions, to improvable reproducibility practices in the area. Many works introducing similarity measures or IC models during the last decade have only implemented or evaluated classic IC-based similarity measures, such as the Resnik [108], Lin [70] and Jiang-Conrath [52] measures, avoiding the replication of IC models and similarity measures introduced by other researchers. Some works have not included all the details of their methods, or the experimental setup to obtain the published results, thus, preventing their reproducibility. Most works have copied results published by others. This latter fact has prevented the invaluable confirmation of previously reported methods and results, which is an essential feature of science. Pedersen [94], and subsequently Fokkens et al. [37], warn of the need to reproduce and validate previous methods and results reported in the literature, a suggestion that we subscribe to in our aforementioned works [56–58], where we also refuted some previous conclusions and warn of finding some contradictory results. A recent study [6,33] on the perception of this reproducibility ‘crisis’ in science shows that the aforementioned reproducibility problems in our area are not the exception but the rule. Precisely, this latter fact has encouraged the recent manifesto for reproducible science [90], which we also subscribe.

And finally, our last motivation is the lack of a gold standard to assist in the exact replication of ontology-based similarity measures and IC models. Most ontology-based similarity measures and intrinsic IC models require the computation of different taxonomical features, such as node depths, hyponym sets, node subsumers, the Least Common Subsumer (LCS), and subsumed leaves, among others. WordNet is a taxonomy with multiple inheritance, thus, some of these features are ambiguously defined, or their computation could be prone to errors. For example, the node depth can be defined as the length of the shortest ascending path from the node to the root, or the length of the longest ascending path as defined by Taieb et al. [43]. Different definitions of depth also lead us to different values for the LCS concepts. On the other hand, the computation of the hyponym set, subsumed leaves and subsumer set requires a careful counting process to avoid node repetitions, as is already noted in [119, §3]. Another potential source of error is the ambiguity in the definition and notation of some IC models and similarity measures. For example, Zhou et al. [134] define the root depth as 1, whilst the standard convention in graph theory is 0. Most authors define the hyponym set as the descendant node set without including the base node itself. However, in [43], the hyponym set also includes the base concept. In addition, we

find works that do not detail the IC models used in their experiments, or how these IC models were built. Finally, many recent hybrid-type measures also require the computation of the length of the shortest path between concepts. These sources of ambiguity and difficulty demand a lot of attention to the fine details for replicating most IC models and similarity measures in the literature. In a recent work [57], we find some contradictory results and difficulties in replicating previous methods and experiments reported in the literature. These reproducibility problems were confirmed in another subsequent work, such as [56], whilst new contradictory results are reported in [58]. Several replication problems were solved with the kind support of most authors. However, we were not able to confirm all previous results, whilst others could not be reproduced through lack of information. As we have explained above, many taxonomical features are ambiguously defined or prone to errors. Thus, all the aforementioned facts lead us to conclude that the exact replication of ontology-based similarity measures and IC models is a hard task, and not exempt from risk. Therefore, it follows that it is urgent and desirable to set off a gold standard for this taxonomical information in order to support the exact replication of the methods reported in the literature.

1.2. Definition of the problem and contributions

This work tackles the problem of designing a scalable and efficient new representation model for taxonomies and a new semantic measures library based on the former, as well as the lack of self-contained reproducible experiments on WordNet-based similarity, tools and resources to assist in the exact replication of methods and experiments previously reported in the literature. In order to bridge the aforementioned gap, the main contributions of this work are as follows: (1) a new and efficient representation model for taxonomies, called *PosetHERep*, which is an adaptation of the half-edge data structure commonly used to represent discrete manifolds and planar graphs in computational geometry; (2) a new Java software library called *Half-Edge Semantic Measures Library (HESML)* based on *PosetHERep*, which implements most ontology-based semantic similarity measures and Information Content (IC) models reported in the literature; (3) a set of reproducible experiments on word similarity based on *HESML* and *ReproZip* [27] with the aim of exactly reproducing the experimental surveys reported in [56–58]; (4) a replication framework and dataset, called *WNSimRep v1*, which is provided as supplementary material at [63], and whose aim is to assist the exact replication of most methods reported in the literature; and finally, (5) the definition and evaluation of a set of scalability and performance benchmarks to compare the state-of-the-art semantic measures libraries.

The rest of the paper is structured as follows. [Section 2](#) introduces the related work. [Section 3](#) introduces the *HESML* software library and the *PosetHERep* representation model for taxonomies. [Section 4](#) introduces a set of reproducible experiments as a companion work to the aforementioned works introduced by Lastra-Díaz and García-Serrano [56–58]. [Section 5](#) briefly introduces the *WNSimRep v1* dataset, which is detailed and made publicly available in [63] as complementary material. [Section 6](#) introduces a series of benchmarks between *HESML* and two state-of-the-art semantic measures libraries with the aim of evaluating and comparing their scalability and performance. [Section 7](#) introduces our discussion of the experimental results. [Section 8](#) introduces our conclusions and future work, whilst [Section 9](#) introduces the revision comments made by the reviewers. Finally, [Appendix A](#) details the resources and datasets included in the *HESML V1R2* distribution.

2. Related work

This section is divided into four subsections according to the categorization of the related work detailed as follows.

[Section 2.1](#) categorizes the family of ontology-based similarity measures. [Section 2.2](#) introduces the IC models which have been implemented in *HESML*. [Section 2.3](#) introduces the main software libraries of ontology-based semantic similarity measures on WordNet reported in the literature. And finally, [Section 2.4](#) introduces some potential applications in information systems. We only introduce herein a categorization of the methods reported in the literature, mainly those implemented in *HESML*. However, for an in-depth review of the latter topics, we refer the reader to the reviews by Lastra-Díaz and García-Serrano on IC-based similarity measures [57] and IC models [56,58], as well as the short review by Batet and Sánchez [12] and the book by Harispe et al. [49].

2.1. Ontology-based semantic similarity measures

[Table 1](#) shows our categorization of the current ontology-based semantic similarity measures into four subfamilies as follows. First, edge-counting measures, the so-called path-based measures, whose core idea is the use of the length of the shortest path between concepts as an estimation of their degree of similarity, such as the pioneering work of Rada et al. [107]. Second, the family of IC-based similarity measures, whose core idea is the use of an Information Content (IC) model, such as the pioneering work of Resnik [108], and the subsequent measures introduced by Jiang and Conrath [52] and Lin [70]. Third, the family of feature-based similarity measures, whose core idea is the use of set-theory operators between the feature sets of the concepts, such as the pioneering work of Tversky [124]. And fourth, other similarity measures that cannot be directly categorized into any previous family, which are based on similarity graphs derived from WordNet [122], novel contributions of the hyponym set [43], or aggregations of other measures [75].

In turn, the more recent IC-based measures can be divided into four subgroups: (1) a first group made up by the aforementioned three classic IC-based similarity measures by Resnik [108], Jiang and Conrath [52], and Lin [70]; (2) a second group defined by those measures that make up an IC model with any function based on the length of the shortest path between concepts, such as the pioneering work of Li et al. [69], and other subsequent works shown in [Table 1](#); (3) a third group of IC-based measures based on the reformulation of different approaches, such as the IC-based reformulations of the Tversky measure by Pirró and Seco [103], and the IC-based reformulation of most edge-counting methods introduced by Sánchez et al. [112]; and finally, (4) a fourth group of IC-based measures based on a monotone transformation of any classic IC-based similarity measure, such as the exponential-like scaling of the Lin measure introduced by Meng and Gu [81], the reciprocal similarity measure of the Jiang-Conrath distance introduced by Garla and Brandt [39], another exponential-like normalization of the Jiang-Conrath distance introduced by Lastra-Díaz and García-Serrano [57], and the monotone transformation of the Lin measure called *FaTH* introduced by Pirró and Euzenat [104]. [Table 2](#) shows a summary of the ontology-based semantic similarity measures implemented by the main publicly available semantic measures libraries.

Finally, we mention five significant further lines of research into ontology-based similarity measures. Stanchev [122] introduces an asymmetric similarity weighted graph derived from WordNet, whilst Martínez-Gil [75] proposes an aggregated similarity measure based on a combination of multiple ontology-based similarity measures and Van Miltenburg [125] proposes a method to compute the semantic similarity between adjectives based on the use of the similarity between their sets of derivational source names in WordNet. More recently, Meymandpour et al. [85] propose several semantic similarity measures for Linked Open Data (LOD) based on IC models, whilst Batet and Sánchez [13] propose a semantic

Table 1

Categorization of the main ontology-based semantic similarity measures based on WordNet reported in the literature and implemented in HESML, excepting those measures with an asterisk (*). The categorization above excludes most GO-based semantic similarity measures, which are in-depth analyzed in a recent survey by Mazandu et al. [76].

Path-based measures	{ Rada et al. [107], Wu & Palmer [130] Leacock & Chodorow [65], Hirst & St-Onge [51]* Pedersen et al. [98], Al-Mubaid & NGuyen [3]
IC-based measures	Classic IC-based measures { Resnik [108] Jiang & Conrath [52] Lin [70]
	Hybrid (path-based) IC-based measures { Li et al. [69] Zhou et al. [133] Meng et al. [83] Gao et al. [38] Lastra-Díaz & García-Serrano(cosw)&C [57]
	Reformulations of other types of measure { Pirró & Seco [103] Sánchez et al. [112]*
	Monotone transformations of classic IC-based measures { Pirró & Euzenat [104] Meng & Gu [81] Garla & Brandt [39] Lastra-Díaz & García-Serrano(cos)&C [57]
Feature-based measures	{ Tversky [124] Batet et al. [14] Sánchez et al. [115]
Other types of measure	{ - Taxonomical features (hyponym sets): Taieb et al. [43] - Aggregation of different of measures: Martínez-Gil [75]* - Asymmetrically weighted graphs based on WordNet: Stanchev [122]* - IC-based reformulation on LinkedOpenData (LOD): Meymandpour et al. [85]* - IC-based reformulation on Wikipedia: Jiang et al. [53]*

relatedness measure based on the combination of highly-accurate ontology-based semantic similarity measures with a resemblance measure derived from corpus statistics.

2.2. Information Content models

The first known IC model is based on corpus statistics and was introduced by Resnik [108], and subsequently detailed in [109]. The main drawback of the corpus-based IC models is the difficulty in getting a well-balanced and disambiguated corpus for the estimation of the concept probabilities. To bridge this gap, Seco et al. [119] introduce the first intrinsic IC model in the literature, whose core hypothesis is that the IC models can be directly computed from intrinsic taxonomical features. Thus, the development of new intrinsic IC-based similarity measures is divided into two subproblems: (1) the proposal of new intrinsic IC models, and (2) the proposal for new IC-based similarity measures. During the last decade, the development of intrinsic IC models has become one of the mainstreams of research in the area. Among the main intrinsic and corpus-based IC models proposed in the literature, we find the proposals by Zhou et al. [133], Sebti and Barfroush [118], Blanchard et al. [18], Sánchez et al. [113,114], Meng et al. [82], Yuan et al. [131], Hadj Taieb et al. [42], Lastra-Díaz and García-Serrano [56,58], Adhikari et al. [1], Aouicha et al. [4,16], and Harispe et al. [46].

Finally, in another recent work, Jiang et al. [53] introduce a new intrinsic IC model based on the Wikipedia category structure which has obtained outstanding results in several word-similarity benchmarks. Table 3 shows a summary of the IC models implemented by the current semantic measures libraries.

2.3. Ontology-based semantic measures libraries

The main publicly available software libraries focusing on the implementation of ontology-based similarity measures based on WordNet are WordNet::Similarity (WNSim) [99] and WS4J [121], whose development is more stable, and the Semantic Measures Library (SML) [47] and the recent WNetSS [15] which are active ongoing projects.

The pioneering WNSim library was developed in Perl by Pedersen et al. [99], and subsequently migrated to Java by Tedeki Shima, under the name of WS4J [121]. WS4J includes, like its parent library, the most significant path-based similarity measures, the three aforementioned classic IC-based measures and several corpus-based IC models [95]. However, WNSim and WS4J do not include most ontology-based similarity measures developed during the last decade, nor any intrinsic IC model. WNSim has been used in a series of papers on word similarity by Patwardhan and Pedersen [93,96], and it has been extended in order to support the UMLS biomedical ontology, thus becoming an independent Perl software library called UMLS::Similarity [78], which is used in a WSD evaluation by McInnes et al. [77]. On the other hand, Harispe et al. [47] introduce the aforementioned SML library, which is the largest semantic measures library. SML is an ongoing project whose v0.9 version implements most classic path-based and IC-based similarity measures as well as several intrinsic IC models; however, it does not include most ontology-based similarity measures and intrinsic IC models developed during the last decade, as shown in Tables 2 and 3. However, SML includes direct support to import OWL and other significant biomedical ontologies such as GO, MeSH and SNOMED-CT. In addition, SML includes several most significant groupwise and pairwise GO-based semantic similarity measures, as

Table 2

Ontology-based semantic similarity measures implemented by the main publicly available software libraries based on WordNet.

Gloss-based similarity measures	WNSim	WS4J	SML	WNetSS	HESML
Banerjee and Pedersen (2003) [7]	X	X			
Patwardhan and Pedersen (2006) [93]	X	X			
Path-based and taxonomy-based measures	WNSim	SML	SML	WNetSS	HESML
Rada et al (1989) [107]	X	X	X	X	X
Wu and Palmer (1994) [130]	X	X	X	X	X
Hirst and St. Onge (1998) [51]	X	X			
Leacock and Chodorow (1998) [65]	X	X		X	X
Stojanovic et al. (2001) [123]			X		
Pekar and Staab (2002) [100]			X		
Li et al (2003) [69], strategy 3				X	X
Li et al (2003) [69], strategy 4					X
Liu et al. (2007) [71]				X	
Pedersen et al (2007) [98]					X
Al-Mubaid and NGuyen (2009) [3]				X	X
Kyogoku et al. (2011) [54]			X		
Hao et al. (2011) [45]				X	
Hadj Taieb et al (2014) [43], sim1				X	X
Hadj Taieb et al (2014) [43], sim2				X	X
IC-based similarity measures	WNSim	WS4J	SML	WNetSS	HESML
Resnik (1995) [108]	X	X	X	X	X
Jiang and Conrath (1997) [52]	X	X	X	X	X
Lin (1998) [70]	X	X	X	X	X
Li et al (2003) strategy 9 [69]					X
Schlicker et al. [116] (GO-based)			X		
Zhou et al (2008) [134]				X	X
Pirró and Seco (2008) [105]				X	X
Pirró and Euzenat (2010) [104], FaITH					X
Garla and Brandt (2012) [39]					X
Meng and Gu (2012) [81]				X	X
Meng et al (2014) [83]					X
Gao et al (2015) [38], strategy 3				X	X
Lastra and García (2015) [57], weighted J&C					X
Lastra and García (2015) [57], cos J&C					X
Lastra and García (2015) [57], cosw J&C					X
Feature-based similarity measures	WNSim	WS4J	SML	WNetSS	HESML
Tversky (1977) [124]				X	
Rodríguez and Egenhofer (2003) [110]				X	
Petrakis et al. (2006) [102]				X	
Sánchez et al (2012) [115]					X

well as a well-supported website and community forum. Thus, SML is currently the most complete and versatile software library reported in the literature. However, there are many other libraries and tools exclusively focused on Gene Ontology (GO), as detailed by Mazandu et al. [76], which should be considered in this specific domain. In addition to the aforementioned Tables 2 and 3, which summarize the methods implemented by the software libraries analyzed herein, Table 4 compares the programming languages and ontologies supported by them.

Finally, we have the WNetSS semantic measures library introduced recently by Aouicha et al. [15], which is based on an off-line pre-processing and caching in a MySQL server of WordNet, as well as all WordNet-based topological features and implemented IC models. As we mentioned previously in Section 1.1.1, the caching strategy used by WNetSS severely impacts its performance and scalability. In addition, WNetSS exhibits two other significant extensibility drawbacks which prevent its use for researching and prototyping of new methods, as follows: (1) the current distribution of WNetSS does not include its source files, thus, their architecture, representation model for taxonomies and implementation details are missing; and (2) the current WNetSS version does not allow any type of functional extension, such as including a new taxonomy parser, as well as a new semantic similarity library or IC model. Finally, despite one of the main motivations of WNetSS being to provide a software implementation for the most recent

methods, looking at Tables 2 and 3, you can see that WNetSS [15] neither implements nor cites many recent similarity measures and IC models reported in the literature.

2.4. Potential applications in Information Systems

Another interesting field of application of the family of ontology-based similarity measures is the problem of business process modeling as detailed below. A very old problem in business process management is the construction and analysis of concept maps that model business processes. Mendling et al. [80] study the current practices in the activity labeling of business processes, whilst Dijkman et al. [32] propose a similarity metric between business process models based on an ad-hoc semantic similarity metric between words in the node labels and attributes, as well as the structural similarity encoded by the concept map topology. Likewise, Leopold et al. [68] propose an automatic refactoring method of activity labels in business process modeling based on the automatic recognition of labeling styles, and Leopold et al. [67] propose the inference of suitable names for business process models automatically. Finally, Montani and Leonardi [89] introduce a framework for the retrieval and clustering of process models based on a semantic and structural distance between models. It is clear that a notion of semantic similarity between components of the models underlies most tasks on process modeling in the latter

Table 3

Intrinsic and corpus-based IC models implemented by the main publicly available software libraries based on WordNet. The above list represents, to the best of our knowledge, all IC models reported in the literature. (*) The Aouicha et al. [16] IC model is implemented in HESML; however, this latter IC model has not yet been evaluated because several missing details need to be clarified by the authors, as described in HESML source code [60].

Corpus-based IC models	WNSim	WS4J	WNetSS	WNetSS	HESML
Resnik corpus-based (1995) [108][109]	X	X	X		X
Lastra & García (2015) [56], CPCorpus					X
Lastra & García (2016) [58], CPreCorpus					X
Intrinsic IC models	WNSim	WS4J	SML	WNetSS	HESML
Seco et al (2004) [119]			X	X	X
Blanchard et al (2008) [18], IC _g					X
Zhou et al (2008) [133]			X	X	X
Sebtí and Barfroush (2008) [118]				X	X
Sánchez et al (2011) [114]			X	X	X
Sánchez et al (2012) [113]					X
Meng et al (2012) [82]				X	X
Harispe (2012) [47]			X		X
Yuan et al (2013) [131]					X
Hadj Taieb et al (2014) [42]				X	X
Adhikari et al (2015) [1]					X
Aouicha et al (2016) [4]					X
Aouicha et al (2016) [16]*				X	X
Harispe et al. (2016) [46]					X
Intrinsic IC models for relatedness measures					
Seddiqi and Aono [120]					
Pirró and Euzenat [104]					
IC models introduced by Lastra-Díaz and García-Serrano (2015) [56]					
CondProbHyponyms					X
CondProbUniform					X
CondProbLeaves					X
CondProbCosine					X
CondProbLogistic					X
IC models introduced by Lastra-Díaz and García-Serrano (2016) [58]					
CondProbRefHyponyms					X
CondProbRefUniform					X
CondProbRefLeaves					X
CondProbRefCosine					X
CondProbRefLogistic					X
CondProbCosineLeaves					X
CondProbRefLogisticLeaves					X
CondProbRefLeavesSubsumerRatio					X

Table 4

Further features of the main publicly available semantic software libraries based on WordNet.

Features	WNSim	WS4J	SML	WNetSS	HESML
Programming language	Perl	Java	Java	Java	Java
Source files availability	public	public	public	no	public
Ongoing development	no	no	yes	yes	yes
Supported ontology file formats:	own parser (own) / external parser				
WordNet	own	own	own	ext WNL	own
OWL			own		
GO			own		
MeSH			own		
SNOMED			own		
RDF triples files			own		

semantic-aware applications. Thus, we argue herein that many of these methods could potentially benefit from the use of ontology-based semantic similarity measures.

3. The HESML software library

HESML V1R2 [60] is distributed as a Java class library (*HESML-V1R2.jar*) plus a test driver application (*HESMLclient.jar*), which has been developed using NetBeans 8.0.2 for Windows, although it has been also compiled and evaluated on Linux-based platforms using the corresponding NetBeans versions. *HESML V1R2* is freely

distributed for any non-commercial purpose under a Creative Commons By-NC-SA-4.0 license² recognized by citing the present work, whilst the *commercial use* of the similarity measures introduced in [57], as well as part of the intrinsic IC models introduced in [56] and [58], is protected by a patent application [58]. HESML is currently being evaluated by Castellanos et al. [22] in a taxonomy recovering task from DBpedia based on Formal Concept Analysis (FCA) methods like the proposed ones in [21]. HESML V1R2 significantly improves the performance of the HESML V1R1 version [59] which was released on September 7 2016 with the original submission of this work.

In order to make the experimental work with HESML easier, as well as supporting the reproducible experiments detailed in Section 4, HESML is distributed as a self-contained development and testing platform including the set of complementary resources shown in Table 22 in appendix, which includes three different WordNet³ versions, a WordNet-based frequency file dataset developed by Ted Pedersen [95], and the five most significant word similarity benchmarks. For this reason, any user of HESML must fulfill the licensing terms of these third-party resources by recognizing their authorship accordingly.

² <https://creativecommons.org/licenses/by-nc-sa/4.0/legalcode> .

³ <https://wordnet.princeton.edu/wordnet/license/> .

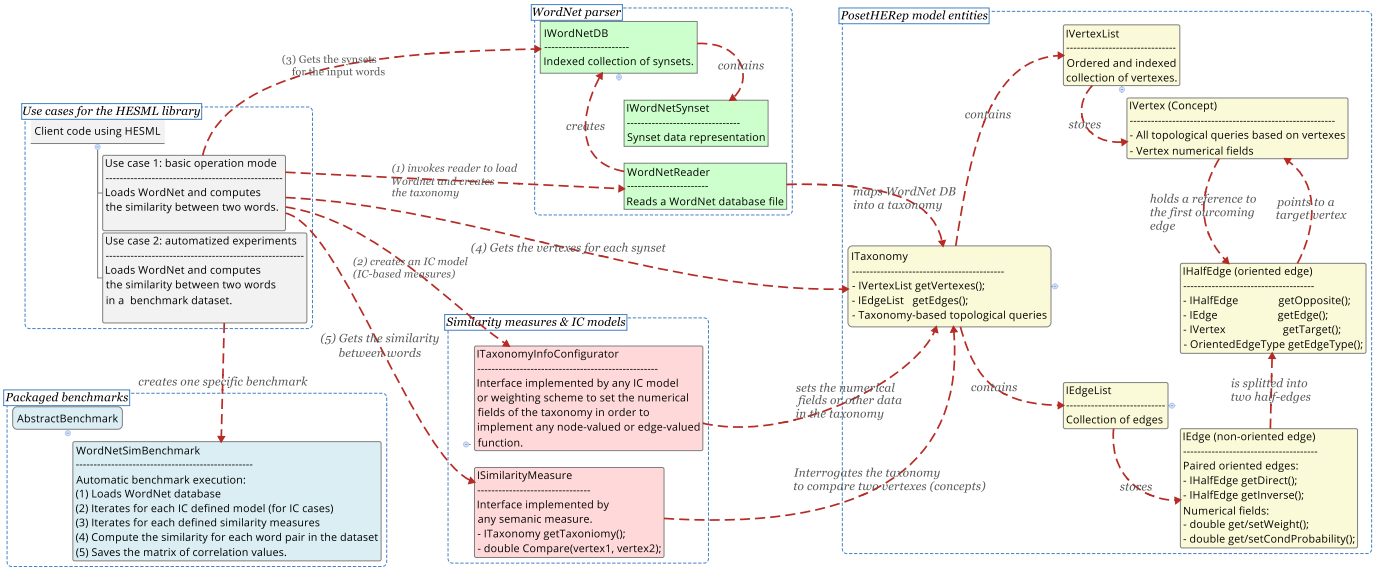


Fig. 1. HESML architecture showing main objects and interfaces. The core HESML component is the half-edge taxonomy representation defined by the yellow entities. Red entities in the block entitled ‘Similarity measures & IC models’ represent the two interfaces that should be implemented to define new IC models and similarity measures. All the HESML objects are provided as Java interfaces, being instantiated by factory objects not represented in the figure. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

HESML V1R2 currently supports the WordNet taxonomy, most ontology-based similarity measures and all the IC models for concept similarity reported in the literature with the only exception of the IC models introduced by Harispe et al. [46], although the latter IC model could be included in future versions. In addition to the aforementioned IC models [46], Seddiqui and Aono [120] and Pirró and Euzenat [104] propose two further intrinsic IC models not implemented by HESML which are based on the integration of all types of taxonomical relationships, and thus especially designed for semantic relatedness measures. In addition, we plan to provide ongoing support for further ontologies such as Wikidata [126] and the Gene Ontology (GO) [5] among others, as well as further similarity and relatedness measures. On the other hand, the HESML architecture allows further similarity measures, IC models and ontology readers to be developed easily. We also urge potential users to propose further functionality. In order to remain up to date on new HESML versions, as well as asking for technical support, we invite the readers to subscribe to the HESML forum detailed in Table 8.

3.1. Software Architecture

The HESML software library is divided into four functional blocks as follows: (1) *PosetHERep* model objects shown in yellow in Fig. 1; (2) abstract interfaces implemented by the IC models or weighting schemes (*ITaxonomyInfoConfigurator*) and all the taxonomy-based similarity measures (*ISimilarityMeasure*) shown in red; (3) ontology readers shown in green; and (4) a family of automatized benchmarks shown in blue, which allow reproducible experiments on ontology-based similarity measures, IC models and word similarity benchmarks with different WordNet versions to be easily implemented, as well as computing and saving the results matrices with Pearson and Spearman correlation values. The automatized benchmarks allow the efficient and exact replication of the experiments and data tables included in the aforementioned works introduced by Lastra-Díaz and García-Serrano. These latter automatized benchmarks can be defined in an XML-based file format, which allows the definition of large experimental surveys without any software coding. All HESML objects are provided as private classes by implementing a set of Java interfaces, thus, they can only be instantiated by invoking the proper factory classes.

All the similarity measures, IC models or weighting schemes are invoked with a reference to the base taxonomy object (*ITaxonomy*) as an input argument, which provides a complete set of queries to retrieve all types of information and topological features. The children, parent, subsumed leaves, ancestor and descendant (hyponym) sets are computed on-the-fly, while the nodes and edges hold the IC values and weights respectively. Any IC model or weighting scheme is defined as an abstract taxonomy processor whose main aim is to annotate the taxonomy with the proper IC values, edge-based weights, concept probabilities or edge-based conditional probabilities. The node-based and edge-based data is subsequently retrieved by the ontology-based semantic similarity measures in their evaluation.

3.2. The *PosetHERep* representation model for taxonomies

PosetHERep is a new and linearly scalable representation model for taxonomies which is introduced herein for the first time. *PosetHERep* is based on our adaptation of the well-known half-edge representation in the field of computational geometry [19], also known as a double-connected edge list [17, § 2.2], in order to efficiently represent and interrogate large taxonomies.

PosetHERep model is the core component of the HESML architecture, it being the mainly responsible for their performance and scalability. Fig. 2 shows the core idea behind the *PosetHERep* representation model: all the outgoing and incoming oriented edges (half-edges) from any vertex are connected in such a way that their connection induces a cyclic ordering on the set of adjacent vertices. Given any single or multiple-root taxonomy $\mathcal{C} = (C, \leq_C)$, we can define its associated graph $G = (V, E)$ in the usual way, in which every concept $c_i \in C$ is mapped onto a vertex $v_i \in C$ and every order relationship between a parent concept and their children is mapped onto an oriented edge, hereinafter called as a half-edge. The core component of the *PosetHERep* model is the *neighbourhood iteration loop* algorithm detailed in Table 5 and three half-edge-valued functions as follows: (1) the *Target* function returns the vertex which the oriented edge points, (2) the *Next* function returns the next outgoing half-edge for each incoming half-edge to any base vertex, and (3) the *Opposite* function returns the opposite and paired half-edge. *PosetHERep* is based on the following *topological*

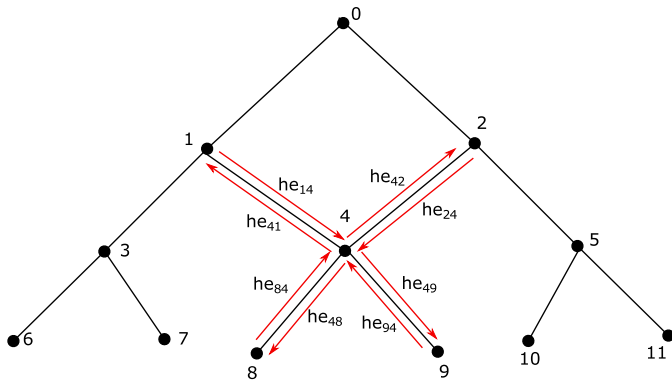


Fig. 2. *PosetHERep*: half-edge representation around the vertex (concept) with id = 4. Every edge is split into two paired and opposite oriented (half) edges. Given the first outgoing half-edge he_{ab} from any vertex a , the set of adjacent vertexes is recovered in linear time through a cyclic iteration, as described by Algorithm 1.

Table 5
Iteration loop from a base vertex in order to recover its adjacent vertexes.

Algorithm 1	Neighbourhood iteration loop
Input:	a base vertex v
Output:	an ordered list $adjVertexes$ of adjacent vertexes
1:	$IVertexList\ adjVertexes$;
2:	$IHalfEdge\ loop = v.firstOutComingEdge$;
3:	do
4:	{
5:	$adjVertexes.Add(loop.Target)$;
6:	$loop = loop.Opposite.Next$;
7:	} while ($loop \neq v.firstOutComingEdge$);

consistency axiom: all the incoming and outgoing half-edges of any vertex are connected in such a way that a full cycle of the *neighbourhood iteration loop* returns the set of adjacency vertexes on any taxonomy vertex. The *HESML* method that inserts the vertexes onto the taxonomy is mainly responsible for the verification of the latter axiom.

The *PosetHERep* model allows most topological queries to be answered in linear time, providing a very efficient implementation for all the graph-traversing algorithms, such as the computation of the depth of the vertexes, ancestor and descendant sets, subsumed leaf sets, and the length of the shortest path between vertexes, among others. Given any taxonomy with an associated graph $G = (V, E)$, it is easy to prove that the memory cost of its *HESML* representation is $O(k_1|V| + k_2|E|)$, in which the constants k_1 and k_2 are defined by the memory size of the vertex and edge attributes. Thus, in any large taxonomy with a small number of concepts with multiple parents we can assume $|V| \approx |E|$, which proves that *HESML* linearly scales with the number of concepts in the taxonomy.

Finally, in order to implement the *PosetHERep* representation model, you must define the behaviour and interface of the six objects shown in yellow in Fig. 1 (*ITaxonomy*, *IVertex*, *IHalfEdge*, *IEdge*, *IVertexList*, and *IEdgeList*), as well as the collection of eight algorithms introduced below. Because of the lack of space, we do not detail seven of these algorithms, thus, we refer the reader to the source code implementing them. The eight algorithms run in linear time as regards the size of the taxonomy, with the only exception being the shortest path algorithm 6. Apart from the output data structures filled by the algorithms detailed below, none of them demands caching or other intensive-memory structures for their implementation. For this reason, the aforementioned algorithms are computationally efficient and scalable.

Algorithm 1. *Neighbourhood iteration loop*. Table 5 details this algorithm, which encodes all the adjacency relationships within the taxonomy. The current *PosetHERep* model only supports ‘is-a’ relationships, because it only supports two types of half-edges: ‘SubClassOf’ and ‘SuperClassOf’. For this reason, the current *HESML* version is only able to represent ‘is-a’ taxonomies. However, the extension of the *PosetHERep* model to manage any type of ontological relationship is straightforward. Thus, we plan to extend its representation capabilities in future versions to include any type of semantic relationship between concepts within an ontology. In addition, *PosetHERep* could be extended to represent many other types of semantic graphs. We also call this algorithm a *vertex iteration loop*, and it is extensively used by most algorithms detailed in this section. Indeed, you can see this piece of code in the software implementation of the aforementioned methods in *HESML*. The iteration loop runs in linear time, it being the time proportional to the number of adjacent vertexes.

Algorithm 2. *Insertion of a vertex in the taxonomy*. This algorithm inserts a new vertex into the taxonomy, as detailed in the source code of the *Taxonomy.addVertex()* function. The method links the vertex to its parent vertexes in order to satisfy the aforementioned *topological consistency axiom*. Once the vertex has been inserted into the taxonomy, it can be directly interrogated without any further inference process, such as that required by other libraries like *SML*. The method runs in linear time, it being the time proportional to the number of adjacent vertexes.

Algorithm 3. *Retrieval of the ancestor set of a vertex*. This algorithm retrieves the ancestor set of any vertex within the taxonomy without caching, as detailed in the source code of the *Vertex.getAncestors()* function. The algorithm climbs up the taxonomy by traversing the ‘SubClassOf’ oriented edges in each local vertex iteration loop. The method runs in linear time, it being the time proportional to the maximum depth of the base vertex.

Algorithm 4. *Retrieval of the descendant set (hyponyms) of a vertex*. This algorithm retrieves the descendant set of any vertex within the taxonomy without caching, as detailed in the source code of the *Vertex.getHyponyms()* function. The algorithm climbs down the taxonomy by traversing the ‘SuperClassOf’ oriented edges in each local vertex iteration loop. The method runs in linear time, it being the time proportional to the difference between the maximum depth of the taxonomy and the base vertex.

Algorithm 5. *Retrieval of the set of subsumed leaves of a vertex*. This algorithm retrieves the set subsumed leaves by any vertex within the taxonomy without caching, as detailed in the source code of the *Vertex.getSubsumedLeaves()* function. The algorithm is identical to the method for retrieving the descendant set with the exception that this method only selects the leaf vertexes, instead of all descendant vertexes. It shares the same computational complexity as algorithm 4.

Algorithm 6. *Shortest path*. This algorithm computes the length of the shortest weighted or unweighted path between two vertexes in the taxonomy, as detailed in the source code of the *Vertex.getShortestPathDistanceTo()* function. The method is a classic Dijkstra algorithm based on a min-priority queue [25,79] and the aforementioned *PosetHERep* vertex iteration loop in order to efficiently traverse the graph. Despite our implementation of the Dijkstra algorithm being very efficient in comparison with other semantic measures libraries, it is still a general-graph method approach with an exponential time complexity.

Algorithm 7. Minimum depth computation. This algorithm computes the minimum depth of the vertex, which is defined as the length of the shortest ascending path from the vertex to the root, as detailed in the source code of the *Vertex.computeMinDepth()* function. The algorithm is divided into two steps: (1) retrieval of the ancestor set, and (2) computation of the shortest ascending path using a modified Dijkstra algorithm constrained to the ancestor set. The core idea of speeding up this algorithm is to reduce the search space for the shortest path algorithm to the ancestor set, which is very efficiently retrieved using algorithm 3. The method runs in linear time, it being the time proportional to the maximum depth of the base vertex.

Algorithm 8. Maximum depth computation. This algorithm computes the maximum depth of the vertex, which is defined as the length of the longest ascending path from the vertex to the root, as detailed in the source code of the *Vertex.computeMaxDepth()* function. This algorithm is identical to the algorithm 7, but in this case it computes the longest ascending path from the vertex to the root.

3.3. Software Functionalities

HESML V1R2 includes the implementation of all the ontology-based similarity measures shown in Table 2, all the IC models shown in Table 3, a set of automatized benchmarks and a reader of WordNet databases. The set of IC models included in *HESML* represents most known intrinsic and corpus-based IC models based on WordNet reported in the literature. The library includes its own WordNet parser and in-memory database representation, it being fully independent of any other software library. In addition, *HESML* defines the *AbstractBenchmark* and *WordnetSimBenchmark* classes in order to provide a family of automatized word similarity benchmarks based on WordNet, as well as an input XML-based reproducible experiment file format which allows all the reproducible experiments detailed in Section 4 and the *WNSimRep v1* dataset to be easily replicated with no software coding.

3.4. Impact

In addition to providing a larger collection of ontology-based similarity measures and intrinsic IC models than other publicly available software libraries, *HESML* provides a more efficient and scalable representation of taxonomies for the prototyping, development and evaluation of ontology-based similarity measures. These aforementioned features convert *HESML* into an open platform to assist the research activities in the area, such as: (1) the development of large experimental surveys, (2) the fast prototyping and development of new methods and applications, (3) the replication of previous methods and results reported in the literature such as in this work, and (4) the dissemination and teaching of ontology-based similarity measures and IC models.

The functionality and software architecture of *HESML* allow the efficient and practical evaluation of large word similarity benchmarks such as SimLex [50] and ontology-based similarity measures based on the length of the shortest path, whose implementation in other software libraries requires a high computational cost that prevents their evaluation in large experimental surveys [58] and datasets. Thus, *HESML* is an essential tool for allowing the fast prototyping and evaluation of new path-based similarity measures on weighted taxonomies or other complex taxonomical features, such as the measures introduced in [57].

Lastra-Díaz and García-Serrano are currently carrying-out a very active research campaign into ontology-based similarity measures and IC models based on *HESML*. Thus, it is expected that *HESML* functionality will grow accordingly. Finally, because of the growing

interest in the integration of ontology-based similarity measures in many applications in the fields of NLP, IR, the Semantic Web and bioengineering, especially genomics, we expect that *HESML* will be helpful and interesting to a larger audience.

3.5. Illustrative examples of use

The *HESMLclient.java* source code file includes a set of sample functions in order to show the functionality of the library as shown in Table 6, which are listed in the function *SampleExperiments()*. All source files are well documented and extensively commented on, in addition to providing a Javadoc documentation. Thus, we think that a careful reading of the source code examples, as well as the understanding of the software architecture detailed in Fig. 1 and the extensibility procedures detailed in Section 3.6, should be enough to use *HESML* to its best advantage. Next, we highlight two examples of use of *HESML*, whilst the next subsection explains how to extend the functionality of the library:

- *Reproducing previous methods and experiments.* We refer the reader to the sample functions in Table 6.
- *Running large experimental surveys.* In addition to checking the aforementioned sample functions, we refer the reader to the Section 4 in which a set of large reproducible experiments is detailed.

3.6. Extending the library

One of the main goals of *HESML* is to replicate previous methods, as well as facilitating the prototyping and development of new methods. The main extensibility axes of the library are the development of new similarity measures and IC models, as well as further ontology parsers. We detail how to carry-out these functionality extensions as follows:

- *Developing and prototyping a new similarity measure.* In order to design a new ontology-based similarity measure, the users must create and register a new class by implementing the *ISimilarityMeasure* interface. The steps to create a new similarity measure are as follows: (1) create a new measure class in the *hesml/measures/impl* namespace, which extends the *SimilaritySemanticMeasure* abstract class and implements the *ISimilarityMeasure* interface; (2) include a new type of measure in the *SimilarityMeasureType.java* enumeration; and (3) register the creation of the new measure in the *getMeasure()* method implemented by the factory class defined in the *hesml/measures/impl/MeasureFactory.java* source file.
- *Developing and prototyping a new IC model.* In order to design a new intrinsic/corpus-based IC model, the users must create and register a new class implementing the *ITaxonomyInfoConfigurator* interface. The steps to create a new intrinsic IC model are as follows: (1) create a new IC model class in the *hesml/configurators/icmodels* namespace, which extends the *AbstractICmodel* class and implements the *ITaxonomyInfoConfigurator* interface; (2) include a new intrinsic IC model type in the *IntrinsicICModelType.java* | *CorpusBasedICModelType.java* enumerations; and (3) register the creation of the new IC model either the *getIntrinsicICmodel()* or *getCorpusICmodel()* methods implemented by the factory class defined in the *hesml/configurators/icmodels/IntrinsicICFactory.java* source file.
- *Developing a new taxonomy reader.* Any taxonomy reader must be able to read a taxonomy file and return an instance of an *ITaxonomy* object. You can use the implementation of the WordNet reader in the *taxonomyreaders/wordnet/impl* namespace as example.

Table 6
Examples of use included in the HESMLclient.java source code file in order to show the functionality of HESML.

HESMLclient method	Description
testAllSimilarityBenchmarks	Runs different types of word similarity benchmarks.
testMultipleICmodelsMultipleICmeasuresBenchmarks	Runs a cross-evaluation of IC models and IC-based similarity measures.
testSingleNonICbasedMeasure	Runs the evaluation of a single non IC-based similarity measures.
testSingleICSimMeasureMultipleICmodels	Runs the evaluation of a single IC-based similarity measure with multiple intrinsic IC models.
testSingleICSimMeasureSingleICmodel	Runs the evaluation of a single IC-based similarity measure with single intrinsic IC models.
testWordPairSimilarity	Shows the computation of the similarity between two words by using the noun database of WordNet and any similarity measure.
testSingleICmodelMultipleICbasedMeasures	Runs the evaluation of a single intrinsic IC model with multiple IC-based similarity measures.
testCorpusBasedSimilarityBenchmarks	Runs the evaluation of multiple corpus-based IC models with multiple IC-based similarity measures.
buildWNSimRepFiles	Builds the WNSimRep v1 dataset.
createTestTaxonomy	This function shows how to create a tree-like taxonomy with the number of vertexes defined by the input parameter. Thus, it shows what should be done by any new ontology parser in order to populate a HESML taxonomy.

Table 7
Complementary Mendeley datasets published with the current work.

Dataset	Content description
HESML V1R2 distribution package [60]	Java source files and NetBeans projects. WordNet 2.1, 3.0 and 3.1 databases. Pedersen's WordNet-based frequency files. Word similarity benchmarks enumerated in table 1.
WordNet-based word similarity reproducible experiments [64]	A ReProZip reproducible experiment file which allows the experimental surveys on WordNet-based word similarity introduced in [57], [56] and [58] to be reproduced, as well as a Zip file with all the raw output files for an easy verification.
WNSimRep v1 dataset [63]	A framework and replication dataset for ontology-based semantic similarity measures and IC models.
HESML_VS_SML [61]	Set of benchmarks introduced herein which evaluate and compare HESML, SML and WNetSS.

Table 8
Summary of technical and legal information of the HESML software library.

HESML source code data	Description
Current code version.	V1R2
Legal Code License.	Creative Commons By-NC-SA 4.0
Permanent code repository used for this version.	http://dx.doi.org/10.17632/t87s78dg78.2
GitHub repository	https://github.com/jjlastra/HESML.git
Software code languages and tools.	Java 8, Java SE DevKit 8, NetBeans 8.0 or higher
Compilation requirements and operating systems.	Java SE Dev Kit 8, NetBeans 8.0 or higher and any Java-compliant operating system.
Documentation and source code examples	This work and the sample source code in the HESMLclient program.
Community forum for questions.	hesml+subscribe@googlegroups.com , hesml+unsubscribe@googlegroups.com

4. The Reproducible Experiments

The aim of this section is to introduce a set of detailed experimental setups in order to exactly replicate the methods and experiments introduced by Lastra-Díaz and García-Serrano in [56–58], whose contributions were stated in the introduction.

4.1. Experimental setup and complementary datasets

We follow the same experimental setup as that detailed in [56] and [58], including the same datasets, preprocessing steps, evaluation metrics, baselines, management of polysemic words and reporting of the results. All the experiments compute the Pearson and Spearman correlation metrics for a set of ontology-based similarity measures on each word similarity benchmark shown in Table 22, as detailed in [56]. Table 7 details the four complementary Mendeley datasets which are distributed in the current work.

4.2. Obtaining and compiling HESML

Table 8 shows the technical information required to obtain and compile the HESML source code and run the experiments detailed in Table 11. There are two different ways of obtaining the HESML source code: (1) by downloading the current version from the permanent Mendeley Data link [60]; and finally, (2) by downloading it from its GitHub repository detailed in Table 8.

Once the source code package has been downloaded or extracted onto your hard drive, the project will have the following folder structure:

1. *HESML_Library*. The root folder of the project.
2. *HESML_Library\HESML*. This folder is the main software library folder containing the NetBeans project and HESML source code. Below this folder you find the *dist* folder which contains the *HESML-V1R2.jar* distribution file generated during the compilation.
3. *HESML_Library\HESMLclient*. This folder contains the source code of the HESMLclient console application. The main aim of the *HESMLclient.jar* application is to provide a collection of sample functions in order to show the HESML functionality, as well as running the collection of reproducible experiments.
4. *HESML_Library\PedersenICmodels*. This folder contains the full WordNet-InfoContent-3.0 collection of WordNet-based frequency files created by Ted Pedersen [95]. The file names denote the corpus used to build each file. The readme file details the method used to build the frequency files, which is also detailed in [97].
5. *HESML_Library\ReproducibleExperiments*. This folder contains three subfolders with the reproducible experiment files shown in Table 11, as well as a XML-schema file called *WordNet-BasedExperiments.xsd*, which describes the syntax of all XML-based experiment files (*.exp), and the *All_paper_tables.exp* file with the definition of all the reproducible experiments shown in Table 11. All files have been created with the XML Spy editor.

Table 9

Configuration of the computers used to reproduce the accompanying set of reproducible experiments, and their running times on the main reproducibility experiments.

Experimental platform	Operating system	CPU	RAM
Ubuntu-base (2011)	Ubuntu MATE 16.04 LTS	Intel Pentium B950 @ 2.10 GHz	4 Gb
Windows-base (2015)	Windows 8.1x64	Intel Core i7-5500U @ 2.40 GHz	8 Gb

In addition, this folder also contains the *RawOutputFiles* subfolder with all the raw output files shown in Table 11, and the *Post-scripts* folder containing the set of post-processing R scripts detailed in Table 12.

6. *HESML_Library\WN_datasets*. This folder contains a set of '*.csv' data files corresponding to the word similarity benchmarks shown in Table 22.
7. *HESML_Library\WordNet-2.1*. This folder contains the database files of WordNet 2.1.
8. *HESML_Library\WordNet-3.0*. This folder contains the database files of WordNet 3.0.
9. *HESML_Library\WordNet-3.1*. This folder contains the database files of WordNet 3.1.

In order to compile *HESML*, you must follow the following steps:

1. Install Java 8, Java SE Dev Kit 8 and NetBeans 8.0.2 or higher in your workstation.
2. Launch NetBeans IDE and open the *HESML* and *HESMLclient* projects contained in the root folder. NetBeans automatically detects the presence of a *nbproject* subfolder with the project files.
3. Select *HESML* and *HESMLclient* projects in the project treeview respectively. Then, invoke the 'Clean and Build project (Shift + F11)' command in order to compile both projects.

4.3. Running the experiments

Table 11 shows the full collection of reproducible experiment files, as well as the corresponding output files that will be generated in order to reproduce the results reported in [57], [56] and [58] respectively.

There are two ways of running the accompanying reproducible experiments: (1) by compiling *HESML* and running the *HESMLclient* program with any input experiment file shown in Table 11, as detailed in Section 4.3.1; or (2) by running the *HESMLv1r1_reproducible_exps.rpz* reproducible experiment file [64] based on *ReproZip*, as detailed in Section 4.3.4. The name of the reproducible experiment files in Table 11 encodes the name of each corresponding table of results that is obtained as output, thus, the table of results that is reproduced. These experiment files reproduce most results reported in [56–58]. However, there are several summary tables in these aforementioned works that are not directly reproduced from the raw output files, thus, the post-processing of several output files is necessary to obtain these missing tables as detailed in Section 4.3.3.

4.3.1. Running the experiments with *HESMLclient*

Once you have compiled the *HESML* and *HESMLclient* projects as detailed in Section 4.2, you are ready to run the reproducible experiments as detailed below. The original *HESMLclient* source code is defined to fetch the required input files from the folder structure of *HESML*. Thus, you only need to follow the steps below:

1. Open a Linux or Windows command prompt in the *HESML_Library\HESMLclient* directory.
2. Run the following command using any reproducible experiment file shown in Table 11:

Table 10

Running times for the main reproducible experiments.

PC name	EAAI_all_tables	KBS_all_tables	AI_all_tables
Ubuntu-base	13491 min \approx 9.37 days	38 s	16 days
Windows-base	—	25 s	—

```
$prompt:> java -jar dist\HESMLclient.jar ..\ReproducibleExperiments \<anyfile.exp>.
```

3. You must run the latter command for each experiment file defined in the aforementioned tables. Optionally, you can run all the experiments automatically by loading any summary file in step 2 above as follows: (1) *EAAI_all_tables.exp*, (2) *KBS_all_tables.exp*, (3) *AI_all_tables.exp*, or (4) *All_paper_tables.exp*. This latter file contains all the experiments shown in Table 11. Table 10 shows the running times for the latter reproducible experiments on the two experimental platforms detailed in Table 9.

Finally, the *WNSimRepv1* dataset [63] can be computed automatically by running the command in step 4 below. The program automatically creates and stores all *WNSimRepv1* data files in the output directory. If the output directory does not exist then it is automatically created.

```
4. $prompt:> java -jar dist\HESMLclient.jar -WNSimRepV1 <outdir>
```

4.3.2. System requirements and performance evaluation

The reproducible experiments detailed in the previous section have been reproduced by the authors in two different experimental platforms shown in Table 9, which are defined by an old low-end laptop called *Ubuntu-base* and a more recent professional laptop called *Windows-base*. The *Ubuntu-base* workstation sets the minimal system requirements in order to reproduce the experiments detailed in previous section, as well as the *ReproZip* package introduced in Section 4.3.4. Table 10 shows the running times for the main reproducible experiments on the two experimental platforms.

4.3.3. Processing of the result files

The running of each experiment file in Table 11 produces one or two comma-separated files (*.csv) with the values separated by a semicolon. The first column in Table 11 shows the number of the table in which the output data computed by each reproducible experiment file (*.exp) appears. All output files are saved in the same folder as their corresponding input experiment files.

Many output files detailed in Table 11 need certain post-processing in order to match the tables shown in the papers exactly. In order to automate this post-processing, we provide the set of R scripts detailed in Table 12. These scripts take the raw output files generated by the experiments in Table 11 and produce the final assembled tables as shown in [56–58], as well as Figs. 2 and 3 showing the interval significance analysis in [56]. The output files shown in the second column in Table 12 are the only files requiring post-processing, the remaining raw output files match the tables shown in the aforementioned works exactly. In order to run

Table 11

Collection of reproducible experiment files for the data tables reported in [57], [56] and [58]. The first column shows the table corresponding to the data generated in the output file. The column entitled 'Measures' denotes the type of similarity measures evaluated by each experiment. Each reproducible experiment file is defined by a XML-based text file with extension (.exp), which can contain the definition of one or more reproducible experiments. Thus, some experiment files produce one output file whilst others produce two output files that must be merged in order to reproduce the original data tables in the papers exactly. Because of the computational cost of the experiments reported in [58], the experiment files corresponding to the latter work generate a single output file containing the Pearson and Spearman correlation metrics that appear separately in the aforementioned work. Thus, it is necessary to split and arrange the columns of the output data tables in order to reproduce the Pearson and Spearman metrics reported in [58] exactly.

Tables	WN	Datasets	IC models	Measures	Metrics	Reproducible experiment file	Output files
Reproducible experiments for the results reported in [57]							
4	All	All	—	Non IC	Pearson	EAAI_table4_nonICmeasures.exp	EAAI_table4_nonICmeasures.csv
5	2.1	RG65, P&S _{full}	intrinsic	IC-based	Pearson	EAAI_table5_RG65_PS.exp	EAAI_table5_RG65.csv EAAI_table5_PS.csv
6	3.0	RG65	all	IC-based	Pearson	EAAI_table6_RG65.csv	EAAI_table6_RG65.csv
7	3.0	P&S _{full}	all	IC-based	Pearson	EAAI_table7_PS.csv	EAAI_table7_PS.csv
8	3.1	RG65, P&S _{full}	intrinsic	IC-based	Pearson	EAAI_table8_RG65_PS.exp	EAAI_table8_RG65.csv EAAI_table8_PS.csv
All	3.0	All	all	all	Pea/Spea	EAAI_all_tables.exp	All output files above
Reproducible experiments for the results reported in [56]							
6	3.0	All	—	H. Taieb [43]	Pea/Spea	KBS_table6_Taieb.exp	KBS_table6_Taieb.csv
7	3.0	RG65	all	IC-based	Pea/Spea	KBS_table7_RG65.csv	KBS_table7_RG65.csv
8	3.0	MC28	all	IC-based	Pea/Spea	KBS_table8_MC28.exp	KBS_table8_MC28.csv
9	3.0	Agirre201	all	IC-based	Pea/Spea	KBS_table9_Agirre201.exp	KBS_table9_Agirre201.csv
10	3.0	P&S _{full}	all	IC-based	Pea/Spea	KBS_table10_PS.exp	KBS_table10_PS.csv
11	3.0	SimLex665	all	IC-based	Pea/Spea	KBS_table11_SimLex665.exp	KBS_table11_SimLex665.csv
All	3.0	All	all	all	Pea/Spea	KBS_all_tables.exp	All output files above
Reproducible experiments for the results reported in [58]							
12	3.0	All	best	All	Pea/Spea	AI_table12.exp	AI_table12.csv
15,16	3.0	RG65	all	IC-based	Pea/Spea	AI_table15_16_RG65.exp	AI_table15_16_RG65.csv
17,18	3.0	MC28	all	IC-based	Pea/Spea	AI_table17_18_MC28.exp	AI_table17_18_MC28.csv
19,20	3.0	Agirre201	all	IC-based	Pea/Spea	AI_table19_20_Agirre201.exp	AI_table19_20_Agirre201.csv
21,22	3.0	P&S _{full}	all	IC-based	Pea/Spea	AI_table21_22_PS.exp	AI_table21_22_PS.csv
23,24	3.0	SimLex665	all	IC-based	Pea/Spea	AI_table23_24_SimLex665.exp	AI_table23_24_SimLex665.csv
All	3.0	All	all	all	Pea/Spea	AI_all_tables.exp	All output files above

the scripts in Table 12, you need to setup the well-known R statistical program⁴ in your workstation. Once R is installed, you need to install the 'BioPhysConnectoR' package, and follow the steps below:

1. Launch the R program
2. Select the menu option 'File->Open script'. Then, load any R-script file contained in the *HESML_Library\Reproducible Experiments\Post-scripts* folder.
3. Edit the 'inputDir' variable at the beginning of the script in order to match the directory containing the raw output files onto your hard drive.
4. Select the menu option 'Edit->Run all'. The final assembled tables will be saved in the input directory defined above, whilst the figures will be shown within R and saved as independent PDF files.

4.3.4. Running the ReproZip experiments

ReproZip is a virtualization tool introduced by Chirigati et al. [27], whose aim is to warrant the exact replication of experimental results onto a different system from that originally used in their creation. ReproZip captures all the program dependencies and is able to reproduce the packaged experiments on any host platform, regardless of the hardware and software configuration used in their creation. Thus, ReproZip warrants the reproduction of the experiments introduced herein in the long term.

The ReproZip program was used for recording and packaging the running of the *HESMLclient* program with all the reproducible experiments shown in Table 11 in the *HESMLv1r1_reproducible_exps.rpz* file available at [64]. This ReproZip file was generated by running ReproZip on the Ubuntu-base

Table 12

Collection of R scripts in order to assemble several tables as shown in the three aforementioned works by Lastra-Díaz and García-Serrano, whose content is not directly obtained from the experimental raw output files. Load the script files in the same order below.

R script file	Post-processing output files and/or figures	
EAAI_final_tables.r	EAAI_final_table_4.csv	
AI_final_tables.r	AI_final_table_10.csv	AI_final_table_11.csv
	AI_final_table_12.csv	
	AI_final_table_15.csv	AI_final_table_16.csv
	AI_final_table_17.csv	AI_final_table_18.csv
	AI_final_table_19.csv	AI_final_table_20.csv
	AI_final_table_21.csv	AI_final_table_22.csv
	AI_final_table_23.csv	AI_final_table_24.csv
	AI_final_table_25.csv	AI_final_table_26.csv
KBS_final_tables.r	KBS_final_table_4.csv	KBS_final_table_6.csv
	KBS_final_table_6.csv	KBS_figure(2,3).pdf

workstation, which was also used to run ReproUnzip based on Docker as detailed below. In order to set up and run the reproducible experiments introduced herein, you need to use ReproUnzip. ReproUnzip can be used with two different virtualization platforms: (1) Vagrant + VirtualBox, or (2) Docker. For a comparison of these two types of virtualization platform, we refer the reader to the survey introduced by Merkel [84], in which the author introduces Docker and compares it with classic Virtual Machines (VM) such as VirtualBox.

Our preferred ReproUnzip configuration is that based on Docker. For instance, in order to setup ReproUnzip based on Docker for Ubuntu, you should follow the detailed steps shown in Table 13, despite several steps possibly being unnecessary depending on your starting configuration. Once ReproUnzip and Docker have been successfully installed, Table 14 shows the detailed instructions to set up and run the reproducible experiments. Those read-

⁴ <https://www.r-project.org/>.

Table 13
Detailed instructions on installing ReprUnzip with Docker for Ubuntu.

Step	Detailed setup instructions
1	sudo apt-get update
2	sudo apt-get install libffi-dev
3	sudo apt-get install libssl-dev
4	sudo apt-get install openssh
5	sudo apt-get install openssh-server
6	sudo apt-get install libsqlite3-dev
7	sudo apt-get install python-dev
8	sudo pip install reprouzip[all]
9	Docker for Ubuntu setup: follow the detailed instructions at https://docs.docker.com/engine/installation/linux/ubuntu/linux/

Table 14
Detailed instructions on how to reproduce the packaged experiments once ReprUnzip has been installed.

Step	Detailed experiment setup and running instructions
1	Setup the ReprUnzip program onto any supported platform (Linux, Windows and MacOS) as detailed in the ReprZip setup page detailed in table.
2	Download the HESMLv1r1 reproducible exps.rpz from its Mendeley repository [64], as detailed in Table 8.
3	Open a command console in the directory containing the HESMLv1r1_reproducible_exps.rpz file and executes the two commands below: (1) reprouzip docker setup HESMLv1r1_reproducible_exps.rpz docker_folder (2) reprouzip docker run docker_folder

Table 15
The first instruction shows a list with the output files generated by the experiments, whilst the second instruction extracts all the output files from the container and downloads them to the current folder.

Step	Detailed instructions to recover the output files
1	reprouzip showfiles docker_folder
2	sudo reprouzip docker download -all docker_folder

Table 16
Tested software platforms for the reproducible experiments based on ReprZip.

Platform	ReprUnzip configuration	Tested
Ubuntu-base	ReprUnzip based on Docker	Yes
Mac Pro (OS X El Capitan – 10.11.6) with 16 Gb RAM	ReprUnzip based on Vagrant	Yes

ers who prefer to use ReprUnzip with VirtualBox instead of Docker can consult the ReprZip installation page.⁵

The running of the reproducible experiments based on Docker for Ubuntu took around 16 days on the aforementioned Ubuntu-base workstation. Once the running has finished, you should follow the instructions shown in Table 15 to recover the output files from the Docker container, as detailed in Table 11. Finally, Table 16 summarizes the software platforms in which the reproducible experiments [64] have been successfully reproduced.

The old low-end Ubuntu-base workstation with only 4Gb RAM is enough to successfully run the experiments detailed in Table 11. However, we suggest a high-end workstation in order to reduce the overall running time.

5. The WNSimRep v1 dataset

WNSimRep v1 is a replication dataset defined by a collection of intrinsic and corpus-based IC models based on WordNet 3.0, which is enriched with the most common taxonomical features used in

the computation of similarity measures and intrinsic IC models, as well as the similarity values reported by most similarity measures in order to assist the replication of previously reported methods and experiments. The *WNSimRep v1* dataset is part of the experimental data reported in our three aforementioned works [56–58], and it was automatically generated using HESML as detailed in Section 4.3.1.

Despite *WNSimRep v1* being based on WordNet 3.0, the proposed framework could be adapted and extended to any type of base ontology, or intrinsic similarity measure. Because of the lack of space, *WNSimRep v1* is detailed in a complementary paper, which together with the dataset files, is publicly available at [63]. *WNSimRep v1* includes three different types of data files: (1) node-valued IC data files with taxonomical features, (2) edge-valued IC data files with the conditional probability between child and parent concepts, and (3) synset-pair-valued data files with taxonomical features and IC-based similarity measures for the synset pairs derived from the classic RG65 benchmark introduced by [111]. The dataset includes 22 intrinsic IC models, 8 corpus-based IC models based on the Resnik method, 8 corpus-based IC models based on the well-founded *CondProbCorpus* IC model, and 8 corpus-based IC model based on the *CondProbRefCorpus*, which have been evaluated with 22 similarity measures. All the corpus-based IC models are derived from the family of “*add1.dat” WordNet-based frequency files included in the Pedersen dataset [95], which is a dataset of corpus-based files created for a series of papers on similarity measures in WordNet, such as [93] and [96]. The dataset includes all the IC models and similarity measures evaluated in the experimental surveys carried-out in the three aforementioned works by Lastra-Díaz and García-Serrano in [56–58].

6. Evaluation of HESML

The goals of the experiments described in this section are as follows: (1) the experimental evaluation of the PosetHERep representation model and HESML, as well as their comparison with the state-of-the-art semantic measures libraries called SML [48] and WNetSS [15]; (2) a study of the impact of the size of the taxonomy on the performance and scalability of the state-of-the-art semantic measures libraries; and finally, (3) the confirmation or refutation of our main hypothesis and research questions; Q1 and Q2 introduced in Section 1.1.

6.1. Experimental setup

Our experiments compare the performance of the HESML V1R2 library version available at [60], with the SML 0.9 library version whose source files are available at GitHub,⁶ and the recent WNetSS library.⁷ We used the compiled *slib-dist-0.9-all-jar.jar* file available at the SML web site⁸ for our experiments. As WNetSS is not distributed with its source files, we were not able to carry-out a side-by-side detailed comparison of WNetSS with HESML and SML, as is done between HESML and SML. Thus, we divided our benchmarks into two blocks: (1) a detailed side-by-side comparison between HESML and SML based on the benchmarks detailed in Table 17 ; and (2) a WordNet-based similarity benchmark based on the SimLex665 dataset in order to evaluate the three aforementioned libraries, which is implemented by the *EvaluateWordNetSimilarity-Dataset* functions in the complementary dataset [61].

In order to evaluate HESML and SML, we have carried out a series of benchmarks based on the creation and interrogation of

⁶ <https://github.com/sharispe/slib>.

⁷ <http://wnetss-api.smr-team.org/>.

⁸ <http://www.semantic-measures-library.org/sml/downloads/releases/sml/0.9/slib-dist-0.9-all-jar.jar>.

⁵ <https://reprouzip.readthedocs.io/en/1.0.x/install.html> .

Table 17

Sequence of benchmarks implemented by the *HSMLeTests* and *SMLTests* classes within the *HESML_vs_SML_tests.jar* program. The test functions carry-out the same operations on both software libraries, thus, their results can be compared directly.

Benchmark	Description
overallCreation	This test creates a tree-like taxonomy with a defined number of vertexes in which each vertex has a random number of children nodes (2 to 8),
avgCreation	<i>overallCreation</i> #vertexes
AncDescLea	This test matches the pre-processing made by the SML, and it consists of the computation of the ancestor and descendant sets of each vertex, and the overall leaf set.
avgAncDesLea	<i>AncDescLea</i> #vertexes
overallCaching	This test measures the number of vertexes cached during the execution of the AncDescLea test (SML pre-processing).
avgCaching	<i>overallCaching</i> #vertexes
avgShortestPath	Average computation time of the shortest path (5 samples).
allMinDepth	Overall computation time of minimum depth for all vertexes.
avgMinDepth	<i>allMinDepth</i> #vertexes
allMaxDepth	Overall computation time of the maximum depth for all vertexes.
avgMaxDepth	<i>allMaxDepth</i> #vertexes
avgLCA	Average time to retrieve the LCA vertex (10,000 samples).
avgMICA	Average time to retrieve the MICA vertex (10,000 samples).
avgSubLea	Average time to retrieve the set of subsumed leaves (10,000 samples).

a sequence of randomly created tree-like taxonomies, whose size grows from 20,000 to 1 million vertexes. The benchmarks have been designed with the aim of evaluating a selection of the most significant topological algorithms used by most ontology-based semantic similarity measures and IC models reported in the literature. Table 17 details the set of benchmarks defined to evaluate the performance of HESML and SML. Because of its high computational cost, we limit the evaluation of the shortest path algorithm to taxonomies with up to 50,000 vertexes. On the other hand, in order to evaluate and compare the performance of WNetSS with HESML and SML, we compare the running-time of the three libraries in the evaluation of the Jiang–Conrath similarity measure [52] with the Seco et al. IC model [119] in the SimLex665 dataset [50].

6.2. Reproducing our benchmarks

All benchmarks detailed in Table 17 are implemented on a single Java console program called *HESML_vs_SML_test.jar*, which is publicly available at [61]. The *HESML_vs_SML* program links directly with the *HESML-V1R2.jar*, *slib-dist-0.9-all.jar.jar* and *WNetSS.jar* files containing the latest publicly available software releases of these libraries. The *HESML_vs_SML* dataset contains all source files and the NetBeans project used to create the entire program, including the pre-compiled version with their dependencies in the ‘dist’ subfolder. The *HESML_vs_SML_test/src* folder contains five files as follows: (1) *HESML_vs_SML_test.java* contains the main function; (2) *HESMLtests.java* contains the functions implementing the aforementioned benchmarks on the HESML V1R2 library; whilst (3) *SMLtests.java* contains the same functions as *HESMLtests.java*, but implementing the benchmarks on the SML 0.9 library; and (4) the *WNetSStests.java* contains the function implementing the WordNet-based similarity benchmark; and finally, (5) the *TestResults.java* file implements a class with the aim of collecting all output results in a structured way. In order to reproduce our benchmarks and see the results reported in Tables 20 and 21, and Fig. 3, you should follow the steps detailed in [61].

6.3. Evaluation metrics

The metrics defined for the comparison of the results are the overall and average running time of the operations, measured in microseconds (μsecs), milliseconds (msecs) or seconds (secs), and the increase in memory derived from the caching process. The measurement of the memory use of a Java program is highly influenced by the Java Virtual Machine (JVM) memory allocation and garbage collector policies. Thus, it is very difficult to carry out a set of measurements on memory use which is reliable, stable

and reproducible. For this reason, the metric used for the caching memory is defined by the exact number of vertexes which are stored in the caching structures. Despite not being able to know the exact caching memory allocated in runtime, we know that it is a multiple of the number of cached vertexes, which is defined by the memory size of each vertex (URIs in SML) and the memory required by the data structures used to store them, typically Java HashSets in SML. Finally, the statistical significance of the results between HESML and SML in the benchmarks detailed in Table 17, as well as the results of the WordNet-based similarity benchmark reported in Table 19, is evaluated using the p-values resulting from the t-student test for the difference mean between the two series of average running times considered as two paired samples sets.

6.4. Results

Tables 20 and 21 show the results of the benchmarks between HESML and SML, whilst Fig. 3 shows a graphical comparison of their performance and Table 18 shows the p-values resulting from the comparison of both series of benchmarks. SML runs out of memory on the taxonomy with 1 million of vertexes. For this reason, we only show the results up to 900,000 vertexes. On the other hand, HESML starts to run out of memory for the same Java heap (4Gb) on taxonomies with 10 million of vertexes or more, a fact that you could check by incrementing the size of the taxonomy in the *HESML_vs_SML* main function. Finally, Table 18 shows the p-values of the benchmarks which are computed using a one-sided t-student distribution on two paired sample sets. Our null hypothesis, denoted by H_0 , is that the difference mean in the average performance between HESML and SML is 0, whilst the alternative hypothesis, denoted by H_1 , is that their average performance is different. For a 5% level of significance, it means that if the p-value is greater than 0.05, we must accept the null hypothesis, otherwise we can reject H_0 with an probability of error of less than the p-value.

Table 19 shows the running-time in milliseconds for five evaluations of the Jiang–Conrath similarity measure in the SimLex665 dataset, together with the average running-time for each library on the Windows-based workstation. We evaluate the WordNet-based similarity benchmark five times to allow a statistical significance analysis and produce a more robust estimation.

7. Discussion

HESML V1R2 significantly outperforms SML 0.9 and sets the new state of the art of the problem. Looking at the Tables 20 and 21,

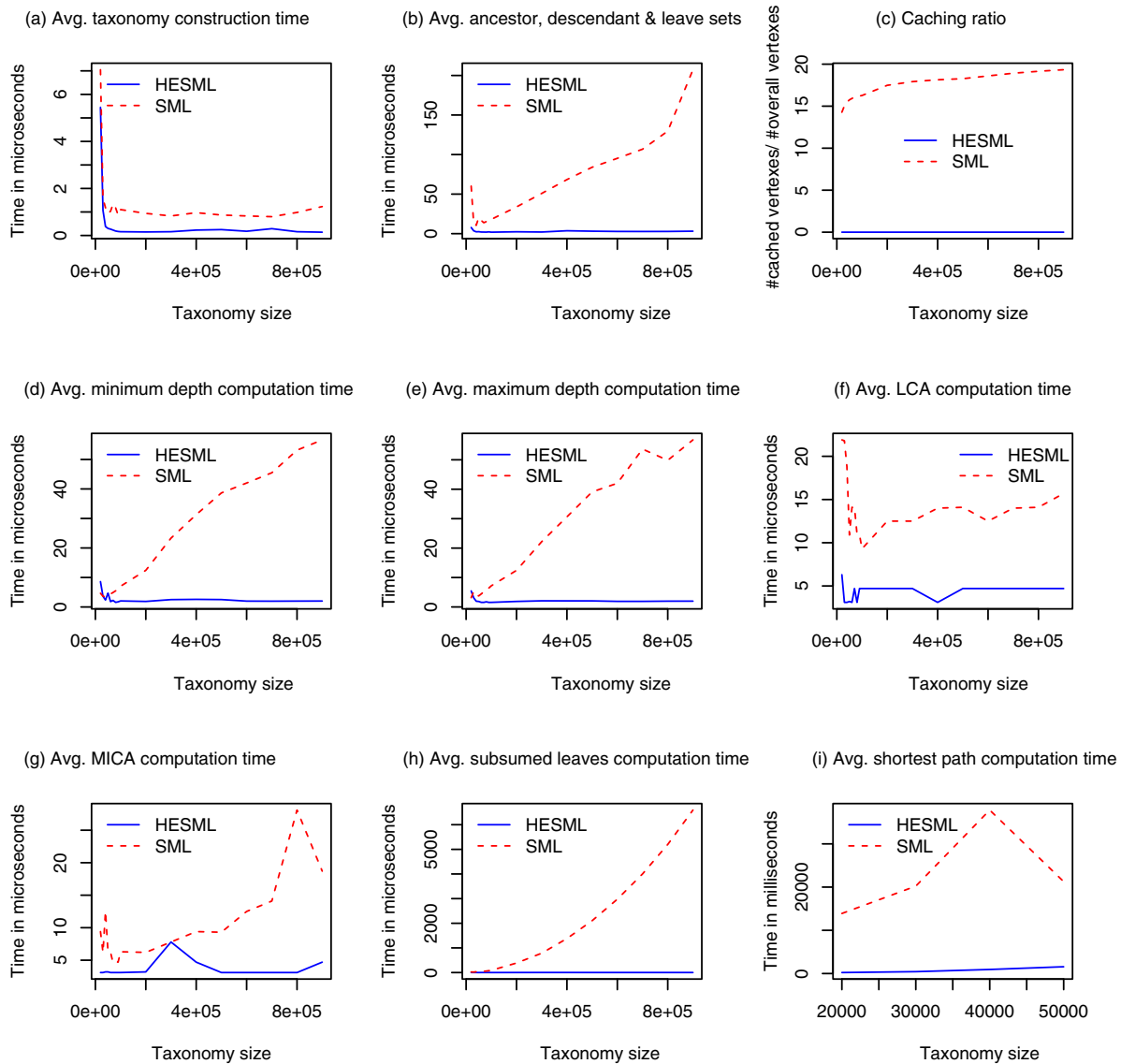


Fig. 3. This figure shows the results obtained by HESML and SML in the series of benchmarks described in the experimental setup, whose values are tabulated in Tables 20 and 21. The computation time is reported in microseconds (μ secs), milliseconds (msecs) or seconds (secs), whilst the increase in memory resulting from the caching carried-out by the SML library is reported in figure(c) as the ratio of the number of cached vertexes as regards the overall number of vertexes, the so called 'taxonomy size'.

Table 18

P-values obtained by using a one-sided t-student distribution for the mean of the differences between two paired samples defined by the HESML and SML benchmark results and a significance level of 95%. The p-values above have been computed by running the *figures_and_table18_Rscript.r* script into the R statistical package, which is provided as complementary material. Any p-value less than 0.05 implies that HESML obtains a statistically significant lower value (running time or caching) than SML. Thus, HESML outperforms SML on this benchmark in a statistically significant manner.

Avg Creation	Avg AncDesLeaves	Avg Caching ratio	Avg Minimum Depth	Avg Maximum Depth	Avg LCA	Avg MICA	Avg Subsumed leaves	Avg shortest path
5.3e-10	4.2e-04	1.6e-18	1.2e-03	8.2e-04	2.3e-09	3.6e-04	6.6e-03	1.0e-02

and Fig. 3, we conclude that HESML outperforms SML in all benchmarks detailed in Table 17. In addition, all p-values in Table 18 are less than 0.05, thus, we conclude that HESML outperforms SML in all benchmarks in a statistically significant manner. Thus, HESML sets the new state of the art in the family of semantic measures libraries in terms of performance and scalability.

Most HESML VIR2 algorithms exhibit linear complexity, thus they are linearly scalable. HESML obtains an almost constant average ratio on most benchmarks, as shown in Tables 20 and 21, and Fig. 3, with the only exception being the shortest path algorithm. The small variation in the average ratios in the aforementioned tables could be attributed to the inherent variability of the time measure-

ment in Java. Thus, most benchmarks exhibit a linear complexity as regards the size of the taxonomy, confirming our theoretical analysis on the scalability of most PosetHERep algorithms introduced in Section 3.2. The set of benchmarks with a constant average ratio, and thus linear complexity, is defined as follows: (1) the creation of the taxonomy (vertex insertion); (2) the retrieval of the ancestor and descendant sets of the vertexes, and the overall leaf set (SML pre-processing); (3) the computation of the minimum and maximum depths of the vertexes; (4) the retrieval of the LCA vertex; (5) the retrieval of the MICA vertex; and (6) the retrieval of the subsumed leaves of the vertexes.

Table 19

Overall running time obtained by the semantic measures libraries in the evaluation of the Jiang-Conrath similarity measure with the Seco et al. IC model in the SimLex665 dataset.

Library	SML	WNetSS	HESML
Run 1 (msecs)	156	177434	110
Run 2 (msecs)	71	177224	89
Run 3 (msecs)	45	177541	97
Run 4 (msecs)	43	173151	85
Run 5 (msecs)	41	179284	82
Avg (msecs)	71.2	176926.8	92.6
t-student p-value (SML, HESML) =			0.147

HESML V1R2 outperforms SML 0.9 including in the benchmarks that use caching. Unlike SML, HESML does not use caching to store any pre-computed set of vertexes. However, HESML significantly outperforms SML in those methods in which SML uses caching, such as the retrieval of the LCA and MICA vertexes, and the set of subsumed leaves of a vertex. On the other hand, HESML makes extensive use of the PosetHeRep model and its algorithms in order to retrieve these objects, outperforming their counterparts based on caching. Thus, our results refute the common belief which states the caching of the entire collection of ancestor and descendant sets is the only solution to speed-up the computation of the aforementioned topological queries. In addition, our results prove that the caching strategy does not only impact the scalability, because of the unneeded and non-linear increment of the memory usage, but also contributes to a low performance as consequence of the continuous interrogations of large hash maps. Specifically, [Table 21](#) shows an almost constant speed-up factor between the average running time for the LCA and MICA benchmarks of HESML as regards SML, which we attribute to the aforementioned interrogations of the caching structures. In the best case, although SML was able to obtain a similar performance to HESML in these tasks after a reengineering of its code, HESML will obtain a better or similar performance without caching. [Table 20](#) shows that SML demands a caching of 19.34 times the taxonomy size for a taxonomy size of 900,000 vertexes, and its caching growing rate is clearly non-linear.

Most SML algorithms exhibit a non-linear time complexity, whilst its best performing methods (LCA and MICA) demand a non-scalable caching strategy. This latter conclusion follows directly from the results shown in [Tables 20](#) and [21](#), as well as the [Fig. 3](#), and our discussion in the previous paragraph.

HESML outperforms most SML benchmarks by several orders of magnitude. As shown in [Tables 20](#) and [21](#), the latter statement is especially significant for large sizes of taxonomy in the following benchmarks: (1) computation of the ancestor and descendant sets, (2) computation of the minimum and maximum depths, (3) computation of the subsumed leaves, and (4) computation of the shortest path between vertexes. SML only obtains good results, for the computation of the MICA and LCA vertexes because of the caching, and even in these two latter cases it is significantly outperformed by HESML. Again, the main problem behind most SML algorithms is its low degree of scalability as consequence of its representation model for taxonomies.

The overall outperformance of HESML on SML proves our main hypothesis and answers our two main research questions positively. Thus, our results allow the following conclusions to be drawn: (1) a new intrinsic representation model for taxonomies as the proposed by PosetHERep is able to improve significantly the performance and scalability of the state-of-the-art semantic measures libraries; and (2) it is possible to significantly improve the performance and scalability of the state-of-the-art semantic mea-

Table 20 Results of the benchmarks between the HESML V1R1 (1.1.01) and SML 0.9 semantic measures libraries.

Taxonomy size	Overall Creation time(msecs)		Avg Creation time(μ secs)		AncDesLeaves time(secs)		Avg AncDesLeaves (μ secs)		Overall caching# Vertexes		Avg Caching		Avg Shortest path(msecs)	
	HESML	SML	HESML	SML	HESML	SML	HESML	SML	HESML	SML	HESML	SML	SML	
20000	109	141	5.45	7.05	0.156	1.203	7.8	60.15	0	285516	0	14.28	228	13894.2
30000	32	47	1.07	1.57	0.109	0.391	3.63	13.03	0	451299	0	15.04	434.4	20218.4
40000	15	47	0.38	1.18	0.094	0.422	2.35	10.55	0	619315	0	15.48	937.6	37819.4
50000	15	47	0.3	0.94	0.125	1.016	2.5	20.32	0	787282	0	15.75	1571.8	21266.2
60000	16	62	0.27	1.03	0.125	1.001	2.08	16.68	0	955252	0	15.92	—	—
70000	16	94	0.23	1.34	0.14	0.969	2	13.84	0	1123293	0	16.05	—	—
80000	15	94	0.19	1.18	0.157	1.219	1.96	15.24	0	1291300	0	16.14	—	—
90000	15	78	0.17	0.87	0.203	1.453	1.453	16.14	0	1459325	0	16.21	—	—
100000	16	110	0.16	1.1	0.187	1.812	1.87	18.12	0	1627322	0	16.27	—	—
200000	31	187	0.15	0.94	0.476	6.766	2.38	33.83	0	3496274	0	17.48	—	—
300000	47	250	0.16	0.83	0.625	15.266	2.08	50.89	0	5376342	0	17.92	—	—
400000	94	390	0.23	0.97	1.469	27.345	3.67	68.36	0	7256275	0	18.14	—	—
500000	125	437	0.25	0.87	1.565	42.001	3.13	84	0	9136247	0	18.27	—	—
600000	110	500	0.18	0.83	1.687	57.143	2.81	95.24	0	11162140	0	18.6	—	—
700000	203	563	0.29	0.8	1.922	74.752	2.75	106.79	0	13242163	0	18.92	—	—
800000	125	781	0.16	0.98	2.25	103.69	2.81	129.61	0	15322176	0	19.15	—	—
900000	125	1110	0.14	1.23	2.829	186.567	3.14	207.3	0	17402182	0	19.34	—	—

Table 21
Results of the benchmarks between the HESML VIRI (1.1.01) and SML 0.9 semantic measures libraries.

Taxonomy size #Vertices	AllMinDepths (msecs)		AvgMinDepth (μ secs)		AllMaxDepth (msecs)		AvgMaxDepth (μ secs)		AvgLCA (μ secs)		AvgMICA (μ secs)		AvgSubLea (μ secs)	
	HESML	SML	HESML	SML	HESML	SML	HESML	SML	HESML	SML	HESML	SML	HESML	SML
20000	172	93	8.6	4.65	109	63	5.45	3.15	6.3	21.9	3.1	9.4	3.2	23.5
30000	109	109	3.63	3.63	94	157	3.13	5.23	3.1	21.8	3.1	6.2	1.6	18.8
40000	94	141	2.35	3.52	78	156	1.95	3.9	3.1	18.8	3.2	12.5	1.5	32.8
50000	235	187	4.7	3.74	93	188	1.86	3.76	3.2	10.9	3.2	6.3	0	29.6
60000	109	265	1.82	4.42	94	266	1.57	4.43	3.1	14.1	3.1	6.2	1.5	56.3
70000	157	344	2.24	4.91	109	359	1.56	5.13	4.7	14.1	3.1	4.6	1.6	50
80000	125	437	1.56	5.46	140	453	1.75	5.66	3.1	11	4.7	1.6	1.6	60.9
90000	157	547	1.74	6.08	140	578	1.56	6.42	4.7	10.9	3.1	4.7	1.6	75
100000	204	703	2.04	7.03	156	719	1.56	7.19	4.7	9.3	3.1	6.3	0	90.6
200000	375	2484	1.88	12.42	375	2485	1.88	12.43	4.7	12.5	3.2	6.2	1.5	386
300000	750	7000	2.5	23.33	625	6656	2.08	22.19	4.7	12.5	7.8	7.8	3.1	785.9
400000	1031	12516	2.58	31.29	828	12250	2.07	30.62	3.1	14	4.7	9.4	3.1	1379.7
500000	1254	19328	2.51	38.66	1031	19547	2.06	39.09	4.7	14.1	3.1	9.3	1.6	2104.8
600000	1188	25203	1.98	42.01	1140	25219	1.9	42.03	4.7	12.5	3.1	12.5	1.6	2976.6
700000	1375	31844	1.96	45.49	1328	37548	1.9	53.64	4.7	14	3.1	14.1	1.6	4001.6
800000	1594	42516	1.99	53.15	1578	39860	1.97	49.83	4.7	14.1	3.1	28.1	1.6	5198.7
900000	1812	50970	2.01	56.63	1781	51095	1.98	56.77	4.7	15.7	4.7	18.7	1.5	6593.9

asures libraries without using any caching strategy by using the PosetHERep model. Likewise, our results confirm our claims in motivation 1.1 in which we state that the caching is a consequence of the use of non-intrinsic naive representation models for taxonomies.

The low performance and scalability of the shortest path algorithm in SML prevents its use in large WordNet-based benchmarks of path-based similarity measures. Looking at Table 20, you can see that SML requires more than 21 s to evaluate the length of the shortest path in a taxonomy with only 50,000 vertexes, it being approximately a half of the WordNet size. This latter fact is especially critical in any WordNet-based word similarity evaluation because the similarity is commonly defined as the maximum similarity in the cartesian product between word senses, thus, it could increase up to two orders of magnitude the latter running time for any path-based similarity measure. On the other hand, looking at Fig. 3.i, you can see the non-linear scaling of the method.

SML obtains the lowest average running-time in the evaluation of a classic IC-based similarity measure in a WordNet-based benchmark, although there is no a statistically significant difference as regard HESML. Looking at Table 19, you can see that SML obtains an average running-time of 71.2 ms, whilst HESML and WNetSS obtain 92.6 and 176,926.8 ms respectively. However, the p-value for the t-student test between SML and HESML is 0.147, thus, there is no a statistically significant difference between these two latter libraries. We attribute this slight advantage of SML on HESML in the WordNet-based test to the WordNet indexing approach of HESML. Despite HESML outperforming SML in the topological algorithms used by the Jiang-Conrath similarity measure, the WordNet indexing and lookup in HESML is up to three times slower than its equivalent in SML. This difference in the performance of the WordNet indexing process between HESML and SML is a consequence of the implementation of two further hashmap lookup operations in HESML, which are not needed by the WordNet indexing approach of SML.

WNetSS obtains the lowest performance in the evaluation of the WordNet-based similarity benchmark, obtaining an average running-time which is more than three orders of magnitude higher than HESML and SML. Table 19 shows that the average running-time of 176,926.8 ms obtained by WNetSS is 2,485 and 1,911 times the average running-time obtained by SML and HESML respectively. This latter fact confirms our statements in Section 1.1.1 on the impact of a software architecture based on a relational database server on the performance and scalability of WNetSS.

Finally, PosetHERep could easily extended in a straightforward way to support any type of semantic relationship, in addition to the 'is-a' taxonomical relationships. Thus, the PosetHERep model could be used as the main building block for large ontologies, and with a proper extension it could be adapted to efficiently manage other non-taxonomical semantic graphs.

7.1. The new state of the art

Our previous discussion allows us to conclude that HESML is the more efficient and scalable semantic measures library between the three libraries evaluated herein. However, there is no a statistically significant difference in the performance of HESML and SML in the evaluation of non path-based similarity measures on WordNet. Thus, SML also provides an efficient and practical solution to evaluate IC-based similarity measures and IC models based on WordNet, despite its performance prevents the evaluation of path-based similarity measures. On the other hand, WNetSS exhibits a poor performance as consequence of its RDBMS-based caching approach, moreover, it does not provide its source files which seriously prevents its evaluation, extensibility and verification. Finally, there would be interesting to carry out a comparison and verifica-

tion of the detailed values reported by each library with the aim of checking and validating their implementation.

8. Conclusions and future work

We have introduced a new and linearly scalable representation model for large taxonomies, called *PosetHERep*, and the *HESML V1R2* [60] semantic measures library based on the former. We have proven in a statistically significant manner that HESML V1R2 is the most efficient and scalable publicly available software library of ontology-based similarity measures and intrinsic IC models based on WordNet. However, there is not a statistically significant difference in the performance of HESML and SML in the evaluation of an IC-based similarity measure based on WordNet, unlike the evaluation of any path-based similarity measure in which HESML is much more efficient. On the other hand, *PosetHERep* and *HESML* have proven, conversely to common belief, that is possible to improve significantly the performance and scalability of the state-of-the-art semantic measures libraries without caching using a proper intrinsic representation model for taxonomies. The performance of WNetSS is more than three orders of magnitude lower than HESML and SML because of its caching strategy based on a relational database.

In addition, we have introduced a set of reproducible experiments based on *ReproZip* [64] and *HESML*, which corresponds to the experimental surveys introduced by Lastra-Díaz and García-Serrano in [57], [56] and [58], as well as the *WNSimRep v1* replication framework and dataset [63] and a benchmark of semantic measures libraries [61].

As forthcoming activities, we plan to extend *HESML* in order to support Wikidata [126] and non “is-a” relationships in the short term, whilst in the mid term, we expect to support the Gene Ontology (GO), MeSH and SNOMED-CT ontologies. In addition, we plan to include further ontology-based similarity measures and IC models reported in the literature, as well as the possibility of importing word embedding files with the aim of allowing the experimental comparison of state-of-the-art ontology-based and corpus-based similarity measures and methods.

9. Revision Comments

This reproducibility paper presents a novel software library (HESML) that implements a plethora of ontology-based semantic similarity measures and information content models. The value of such library is indubitable, since it provides a benchmark to compare existing and potentially new approaches in the field. By using and evaluating the implemented measures and models, researchers are able to thoroughly compare the available implementations and uncover which are the measures that more accurately mimic hu-

man understanding. In addition, because the source code is provided, new models and measures can more easily be built on top of the existing ones, facilitating the progress of the research on similarity measures.

While reviewing this manuscript, a few issues around reproducibility were brought into discussion. One issue was related to post-processing: ideally, for reproducibility purposes, the post-processing of output files should be as automatic as possible to facilitate the generation of the final results and figures of the paper. Evaluating performance and scalability is also key to reproducibility, since this makes the library more appealing for readers and researchers who will use it and perform experiments in potentially different computational platforms. Last, not only the instructions to run the library should be clear, but also the implemented modules and functions should be well described to make the library extendable and more useful. The authors satisfactorily took all our comments into account and significantly improved their artifact. It is worth noting that an important outcome of this submission and the reviews was the improvement in performance and scalability of the library, which will greatly benefit every researcher working in this area.

We would like to thank the authors for providing such a valuable artifact to the community, and for their great effort in making sure that all the instructions for building and using the library are clear, and all the experimental results can be reproduced effortlessly.

Acknowledgements

Ted Pedersen kindly answered all our questions and provided his WordNet-based frequency files dataset, whilst Sébastien Harispe provided the SML source code and his total support in evaluating it. Mohamed Hadj Taieb kindly offered us his total support in replicating their similarity measures exactly. Jorge Martínez-Gil, Emmanuel Pothos, Emiel Van Miltenburg and Lubomir Stanchev kindly answered our questions about their methods. Ángel Castellanos proposed us the use of HESML to manage FCA-based applications. Mark Hallett checked the proper use of English. Finally, we are grateful to the reviewers for their suggestions in order to improve significantly this work. This work has been partially supported by the Spanish Musacces (S2015/HUM3494) and VEMODALEN (TIN2015-71785-R) Projects.

Appendix A. Resources in the HESML distribution

Table 22 details the resources and datasets included in the HESML V1R2 distribution.

Table 22
Collection of resources distributed as supplementary material of the present work and included the HESML V1R2 distribution package.

Reference works	Acronym	Resource type	Licensing type
This work and [60]	HESML V1R2	Java software library	CC By-NC-SA 4.0
This work and [63]	WNSimRep v1	Replication dataset	CC By-NC 3.0
Miller [87], Fellbaum [35]	WordNet 2.1	Ontology-based lexicon	Attribution
Miller [87], Fellbaum [35]	WordNet 3.0	Ontology-based lexicon	Attribution
Miller [87], Fellbaum [35]	WordNet 3.1	Ontology-based lexicon	Attribution
Rubenstein and Goodenough [111]	RG65	Word similarity benchmark	Attribution
Miller and Charles [88]	MC28	Word similarity benchmark	Attribution
Agirre et al. [2]	Agirre201	Word similarity benchmark	Attribution
Pirró [103]	<i>P&S_{full}</i>	Word similarity benchmark	Attribution
Hill et al. [50]	SimLex665	Word similarity benchmark	Attribution
Patwardhan and Pedersen [93], Pedersen [96]	WN-IC-3.0.tar	WN-based frequency files	Attribution

References

- [1] A. Adhikari, S. Singh, A. Dutta, B. Dutta, A novel information theoretic approach for finding semantic similarity in WordNet, in: Proceedings of IEEE International Technical Conference (TENCON-2015), IEEE, Macau, China, 2015, pp. 1–6, doi:10.1109/TENCON.2015.7372780.
- [2] E. Agirre, E. Alfonseca, K. Hall, J. Kravalova, M. Paşca, A. Soroa, A study on similarity and relatedness using distributional and WordNet-based approaches, in: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, in: NAACL '09, Association for Computational Linguistics, Stroudsburg, PA, USA, 2009, pp. 19–27.
- [3] H. Al-Mubaid, H.A. Nguyen, Measuring semantic similarity between biomedical concepts within multiple ontologies, IEEE Trans. Syst. Man Cybern. Part C Appl. Rev. 39 (4) (2009) 389–398, doi:10.1109/TSMCC.2009.2020689.
- [4] M.B. Aouicha, M.A.H. Taieb, Computing semantic similarity between biomedical concepts using new information content approach, J. Biomed. Inf. 59 (2016) 258–275, doi:10.1016/j.jbi.2015.12.007.
- [5] M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J. Michael Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, M.A. Harris, D.P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J.C. Matese, J.E. Richardson, M. Ringwald, G.M. Rubin, G. Sherlock, Gene Ontology: tool for the unification of Biology, Nat. Genet. 25 (1) (2000) 25–29, doi:10.1038/75556.
- [6] M. Baker, 1,500 scientists lift the lid on reproducibility, Nature 533 (7604) (2016) 452–454, doi:10.1038/533452a.
- [7] S. Banerjee, T. Pedersen, Extended gloss overlaps as a measure of semantic relatedness, in: Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence (IJCAI), 2003, pp. 805–810. Acapulco, México.
- [8] R. Banjade, N. Maharjan, N.B. Niraula, V. Rus, D. Gautam, Lemon and tea are not similar: measuring word-to-word similarity by combining different methods, in: A. Gelbukh (Ed.), Proceedings of the 16th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing), LNCS, 9041, Springer, Cairo, Egypt, 2015, pp. 335–346, doi:10.1007/978-3-319-18111-0_25.
- [9] M. Batet, A study on semantic similarity and its application to clustering: enabling the classification of textual data, VDM Verlag, 2011.
- [10] M. Batet, Ontology-based semantic clustering, AI Commun. Eur. J. Artif. Intell. 4 (3) (2011) 291–292, doi:10.3233/AIC-2011-0501.
- [11] M. Batet, A. Erola, D. Sánchez, J. Castellà-Roca, Utility preserving query log anonymization via semantic microaggregation, Inf. Sci. 242 (2013) 49–63, doi:10.1016/j.ins.2013.04.020.
- [12] M. Batet, D. Sánchez, A review on semantic similarity, in: M. Khosrow-Pour (Ed.), Encyclopedia of Information Science and Technology, third edition, ISI Global, 2015, pp. 7575–7583, doi:10.4018/978-1-4666-5888-2.ch746.
- [13] M. Batet, D. Sánchez, Improving semantic relatedness assessments: ontologies meet textual corpora, Procedia Comput. Sci. 96 (2016) 365–374, doi:10.1016/j.procs.2016.08.149.
- [14] M. Batet, D. Sánchez, A. Valls, An ontology-based measure to compute semantic similarity in biomedicine, J. Biomed. Inf. 44 (1) (2011) 118–125, doi:10.1016/j.jbi.2010.09.002.
- [15] M. Ben Aouicha, M.A.H. Taieb, A. Ben Hamadou, SISR: system for integrating semantic relatedness and similarity measures, Soft Comput. (2016) 1–25, doi:10.1007/s00500-016-2438-x. <http://dx.doi.org/10.1007/s00500-016-2438-x>
- [16] M. Ben Aouicha, M.A.H. Taieb, A. Ben Hamadou, Taxonomy-based information content and wordnet-wiktionary-wikipedia glosses for semantic relatedness, Appl. Intell. (2016) 1–37, doi:10.1007/s10489-015-0755-x.
- [17] M. de Berg, M. Van Kreveld, M. Overmars, O. Schwarzköpf, Computational Geometry: Algorithms and Applications, Springer-Verlag, 1997.
- [18] E. Blanchard, M. Harzallah, P. Kuntz, A generic framework for comparing semantic similarities on a subsumption hierarchy, in: M. Ghallab, C.D. Spyropoulos, N. Fakotakis, N. Avouris (Eds.), Proceedings of the ECAI, Frontiers in Artificial Intelligence and Applications, 178, IOS Press, 2008, pp. 20–24, doi:10.3233/978-1-58603-891-5-20.
- [19] M. Botsch, S. Steinberg, S. Bischoff, L. Kobbelt, OpenMesh - a generic and efficient polygon mesh data structure, CiteseerX (2002).
- [20] A. Budanitsky, G. Hirst, Evaluating WordNet-based measures of lexical semantic relatedness, Comput. Ling. 32 (1) (2006) 13–47, doi:10.1162/coli.2006.32.1.13.
- [21] A. Castellanos, J. Cigarrán, A. García-Serrano, Formal concept analysis for topic detection: a clustering quality experimental analysis, Inf. Syst. (2017). <http://dx.doi.org/10.1016/j.is.2017.01.008>.
- [22] A. Castellanos, A. García-Serrano, J. Cigarrán, Linked data-based conceptual modelling for recommendation: a FCA-based approach, in: M. Hepp, Y. Hoffner (Eds.), E-Commerce and Web Technologies, Lecture Notes in Business Information Processing, Springer International Publishing, 2014, pp. 71–76, doi:10.1007/978-3-319-10491-1_8.
- [23] P. Castells, M. Fernández, D. Vallet, An adaptation of the vector-space model for ontology-based information retrieval, IEEE Trans. Knowl. Data Eng. 19 (2) (2007) 261–272.
- [24] J.M. Chaves-González, J. Martínez-Gil, Evolutionary algorithm based on different semantic similarity functions for synonym recognition in the biomedical domain, Knowl.-Based Syst. 37 (2013) 62–69, doi:10.1016/j.knsys.2012.07.005.
- [25] M. Chen, R.A. Chowdhury, V. Ramachandran, D.L. Roche, L. Tong, Priority Queues and Dijkstra's Algorithm, Technical Report TR-07-54, Computer Science Department, University of Texas at Austin, 2007. <http://www3.cs.stonybrook.edu/~rezaul/papers/TR-07-54.pdf>.
- [26] F. Chirigati, R. Capone, R. Rampin, J. Freire, D. Shasha, A collaborative approach to computational reproducibility, Inf. Syst. 59 (2016) 95–97, doi:10.1016/j.is.2016.03.002.
- [27] F. Chirigati, R. Rampin, D. Shasha, J. Freire, ReproZip: computational reproducibility with ease, in: Proceedings of the 2016 ACM SIGMOD International Conference on Management of Data (SIGMOD), 16, bigdata.poly.edu, 2016, pp. 2085–2088.
- [28] F.M. Couto, H.S. Pinto, The next generation of similarity measures that fully explore the semantics in biomedical ontologies, J. Bioinf. Biol. 11 (5) (2013) 1371001, doi:10.1142/S0219720013710017.
- [29] F.M. Couto, M.J. Silva, P.M. Coutinho, Measuring semantic similarity between Gene Ontology terms, Data . Knowl. Eng. 61 (1) (2007) 137–152, doi:10.1016/j.datak.2006.05.003.
- [30] V. Cross, X. Hu, Using semantic similarity in Ontology alignment, in: Proceedings of the Sixth International Workshop on Ontology Matching (OM), 10th Int. Semantic Web Conference (ISWC 2011), 2011, pp. 61–72. Bonn Germany.
- [31] G.G. Dagher, B.C.M. Fung, Subject-based semantic document clustering for digital forensic investigations, Data . Knowl. Eng. 86 (2013) 224–241, doi:10.1016/j.datak.2013.03.005.
- [32] R. Dijkman, M. Dumas, B. van Dongen, R. Käärik, J. Mendling, Similarity of business process models: metrics and evaluation, Inf. Syst. 36 (2) (2011) 498–516, doi:10.1016/j.is.2010.09.006.
- [33] Editorial, Reality check on reproducibility, Nature 533 (7604) (2016) 437, doi:10.1038/533437a.
- [34] J. Fährdrich, S. Weber, S. Ahrndt, Design and use of a semantic similarity measure for interoperability among agents, in: M. Klusch, R. Unland, O. Shehry, A. Pokahr, S. Ahrndt (Eds.), Multiagent System Technologies, Lecture Notes in Computer Science, Springer International Publishing, 2016, pp. 41–57, doi:10.1007/978-3-319-45889-2_4.
- [35] . WordNet: An Electronic Lexical Database, in: C. Fellbaum (Ed.), MIT Press, Cambridge, MA, 1998.
- [36] S. Fernando, M. Stevenson, A semantic similarity approach to paraphrase detection, in: Proceedings of the 11th Annual Research Colloquium of the UK Special-interest group for Computational Linguistics, 2008, pp. 45–52. Oxford, UK.
- [37] A. Fokkens, M. Van Erp, M. Postma, T. Pedersen, P. Vossen, N. Freire, Offspring from reproduction problems: what replication failure teaches us, in: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL, Sofia, Bulgaria, 2013, pp. 1691–1701.
- [38] J.B. Gao, B.W. Zhang, X.H. Chen, A WordNet-based semantic similarity measurement combining edge-counting and information content theory, Eng. Appl. Artif. Intell. 39 (2015) 80–88, doi:10.1016/j.engappai.2014.11.009.
- [39] V.N. Garla, C. Brandt, Semantic similarity in the biomedical domain: an evaluation across knowledge sources, BMC Bioinf. 13:261 (2012), doi:10.1186/1471-2105-13-261.
- [40] T. Grego, F.M. Couto, Enhancement of chemical entity identification in text using semantic similarity validation, PloS One 8 (5) (2013) e62984, doi:10.1371/journal.pone.0062984.
- [41] P.H. Guzzi, M. Mina, C. Guerra, M. Cannataro, Semantic similarity analysis of protein data: assessment with biological features and issues, Briefings Bioinf. 13 (5) (2012) 569–585, doi:10.1093/bib/bbr066.
- [42] M.A. Hadj Taieb, M. Ben Aouicha, A. Ben Hamadou, A new semantic relatedness measurement using WordNet features, Knowl. Inf. Syst. 41 (2) (2014) 467–497, doi:10.1007/s10115-013-0672-4.
- [43] M.A. Hadj Taieb, M. Ben Aouicha, A. Ben Hamadou, Ontology-based approach for measuring semantic similarity, Eng. Appl. Artif. Intell. 36 (2014) 238–261, doi:10.1016/j.engappai.2014.07.015.
- [44] M.A. Hadj Taieb, M. Ben Aouicha, Y. Bourouis, FM3s: features-based measure of sentences semantic similarity, in: E. Onieva, I. Santos, E. Osaba, H. Quintián, E. Corchado (Eds.), Proceedings of the 10th International Conference on Hybrid Artificial Intelligent Systems (HAIS 2015), LNCS, 9121, Springer, Bilbao, Spain, 2015, pp. 515–529, doi:10.1007/978-3-319-19644-2_43.
- [45] D. Hao, W. Zuo, T. Peng, F. He, An approach for calculating semantic similarity between words using WordNet, in: Proceedings of the Second International Conference on Digital Manufacturing Automation, IEEE, 2011, pp. 177–180, doi:10.1109/ICDMA.2011.50.
- [46] S. Harispe, A. Imoussaten, F. Trusset, J. Montmain, On the consideration of a bring-to-mind model for computing the information content of concepts defined into ontologies, in: Proceedings of the IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2015), IEEE, Istanbul, Turkey, 2015, pp. 1–8, doi:10.1109/FUZZ-IEEE.2015.7337964.
- [47] S. Harispe, S. Ranwez, S. Janaqi, J. Montmain, The semantic measures library and toolkit: fast computation of semantic similarity and relatedness using biomedical ontologies, Bioinf. 30 (5) (2014) 740–742, doi:10.1093/bioinformatics/btt581.
- [48] S. Harispe, S. Ranwez, S. Janaqi, J. Montmain, The semantic measures library: assessing semantic similarity from knowledge representation analysis, in: E. Métais, M. Roche, M. Teisseire (Eds.), Proceedings of the 19th International Conference on Applications of Natural Language to Information Systems (NLDB 2014), LNCS, 8455, Springer, Montpellier, France, 2014, pp. 254–257, doi:10.1007/978-3-319-07983-7_37.

- [49] S. Harispe, S. Ranwez, S. Janaqi, J. Montmain, Semantic similarity from natural language and ontology analysis, *Synthesis Lectures on Human Language Technologies*, 8, Morgan & Claypool publishing, 2015, doi:10.2200/S00639ED1V01Y201504HLT027.
- [50] F. Hill, R. Reichart, A. Korhonen, SimLex-999: evaluating semantic models with (Genuine) similarity estimation, *Comput. Ling.* 41 (4) (2015) 665–695, doi:10.1162/COLL_a_00237.
- [51] G. Hirst, D. St-Onge, Lexical chains as representations of context for the detection and correction of malapropisms, in: C. Fellbaum (Ed.), *WordNet: An electronic lexical database*, Massachusetts Institute of Technology, 1998, pp. 305–332.
- [52] J.J. Jiang, D.W. Conrath, Semantic similarity based on corpus statistics and lexical taxonomy, in: *Proceedings of International Conference Research on Computational Linguistics (ROCLING X)*, 1997, pp. 19–33.
- [53] Y. Jiang, W. Bai, X. Zhang, J. Hu, Wikipedia-based information content and semantic similarity computation, *Inf. Process. Manage.* 53 (1) (2017) 248–265, doi:10.1016/j.ipm.2016.09.001.
- [54] R. Kyogoku, R. Fujimoto, T. Ozaki, T. Ohkawa, A method for supporting retrieval of articles on protein structure analysis considering users' intention, *BMC Bioinf.* 12 Suppl 1 (2011) S42, doi:10.1186/1471-2105-12-S1-S42.
- [55] J.J. Lastra-Díaz, *Intrinsic Semantic Spaces for the Representation of Documents and Semantic Annotated Data*, Universidad Nacional de Educación a Distancia (UNED), Department of Computer Languages and Systems, 2014 Master's thesis. <http://e-spacio.uned.es/fez/view/bibliuned:master-ETSIInformatica-LSI-Jlastra>.
- [56] J.J. Lastra-Díaz, A. García-Serrano, A new family of information content models with an experimental survey on WordNet, *Knowl.-Based Syst.* 89 (2015) 509–526, doi:10.1016/j.knsys.2015.08.019.
- [57] J.J. Lastra-Díaz, A. García-Serrano, A novel family of IC-based similarity measures with a detailed experimental survey on WordNet, *Eng. Appl. Artif. Intell.* 46 (2015) 140–153, doi:10.1016/j.engappai.2015.09.006.
- [58] J.J. Lastra-Díaz, A. García-Serrano, A Refinement of the Well-Founded Information Content Models with a very Detailed Experimental Survey on WordNet, Technical Report, NLP and IR Research Group, ETSI Informática, Universidad Nacional de Educación a Distancia (UNED), 2016. <http://e-spacio.uned.es/fez/view/bibliuned:DptoLSI-ETSI-Informes-Jlastra-refinement>.
- [59] J.J. Lastra-Díaz, A. García-Serrano, in: HESML V1R1 Java software library of ontology-based semantic similarity measures and information content models, 2016, doi:10.17632/t87s78dg78.1. (Mendeley Data, v1). <http://dx.doi.org/10.17632/t87s78dg78.1>.
- [60] J.J. Lastra-Díaz, A. García-Serrano, in: HESML V1R2 Java software library of ontology-based semantic similarity measures and information content models, 2016, doi:10.17632/t87s78dg78.2. (Mendeley Data, v2). <http://dx.doi.org/10.17632/t87s78dg78.2>.
- [61] J.J. Lastra-Díaz, A. García-Serrano, in: HESML_vs_SML: scalability and performance benchmarks between the HESML V1R2 and SML 0.9 semantic measures libraries, 2016, doi:10.17632/5hg3z85wf4.1. (Mendeley Data, v1). <http://dx.doi.org/10.17632/5hg3z85wf4.1>.
- [62] J.J. Lastra-Díaz, A. García-Serrano, System and method for the indexing and retrieval of semantically annotated data using an ontology-based information retrieval model, United States Patent and Trademark Office (USPTO) Application, US2016/0179945 A1(2016).
- [63] J.J. Lastra-Díaz, A. García-Serrano, in: WNSimRep: a framework and replication dataset for ontology-based semantic similarity measures and information content models, 2016, doi:10.17632/mpr2m8pycs.1. (Mendeley Data v1). <http://dx.doi.org/10.17632/mpr2m8pycs.1>.
- [64] J.J. Lastra-Díaz, A. García-Serrano, in: WordNet-based word similarity reproducible experiments based on HESML V1R1 and ReprZip, 2016. (Mendeley Data, v1). <http://dx.doi.org/10.17632/65pxgskhz9.1>.
- [65] C. Leacock, M. Chodorow, Combining local context and WordNet similarity for word sense identification, in: C. Fellbaum (Ed.), *WordNet: An electronic lexical database*, MIT Press, 1998, pp. 265–283.
- [66] M.C. Lee, A novel sentence similarity measure for semantic-based expert systems, *Expert Syst. Appl.* 38 (5) (2011) 6392–6399, doi:10.1016/j.eswa.2010.10.043.
- [67] H. Leopold, J. Mendling, H.A. Reijers, M. La Rosa, Simplifying process model abstraction: techniques for generating model names, *Inf. Syst.* 39 (2014) 134–151, doi:10.1016/j.is.2013.06.007.
- [68] H. Leopold, S. Smirnov, J. Mendling, On the refactoring of activity labels in business process models, *Inf. Syst.* 37 (5) (2012) 443–459, doi:10.1016/j.is.2012.01.004.
- [69] Y. Li, Z.A. Bandar, D. McLean, An approach for measuring semantic similarity between words using multiple information sources, *IEEE Trans. Knowl. Data Eng.* 15 (4) (2003) 871–882, doi:10.1109/TKDE.2003.1209005.
- [70] D. Lin, An information-theoretic definition of similarity, in: *Proceedings of the 15th International Conference on Machine Learning*, 98, 1998, pp. 296–304. Madison, WI.
- [71] X.Y. Liu, Y.M. Zhou, R.S. Zheng, Measuring semantic similarity in Wordnet, in: *Proceedings of the 2007 International Conference on Machine Learning and Cybernetics*, 6, IEEE, 2007, pp. 3431–3435, doi:10.1109/ICMLC.2007.4370741.
- [72] P.W. Lord, R.D. Stevens, A. Brass, C.A. Goble, Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation, *Bioinf.* 19 (10) (2003) 1275–1283, doi:10.1093/bioinformatics/btg153.
- [73] F. Mandreoli, R. Martoglia, Knowledge-based sense disambiguation (almost) for all structures, *Inf. Syst.* 36 (2) (2011) 406–430, doi:10.1016/j.is.2010.08.004.
- [74] S. Martínez, D. Sánchez, A. Valls, Ontology-based anonymization of categorical values, in: *Modeling Decisions for Artificial Intelligence*, in: LNCS, 6408, Springer Berlin Heidelberg, 2010, pp. 243–254, doi:10.1007/978-3-642-16292-3_24.
- [75] J. Martínez-Gil, CoTO: a novel approach for fuzzy aggregation of semantic similarity measures, *Cognit. Syst. Res.* 40 (2016) 8–17, doi:10.1016/j.cogsys.2016.01.001.
- [76] G.K. Mazandu, E.R. Chimusa, N.J. Mulder, Gene Ontology semantic similarity tools: survey on features and challenges for biological knowledge discovery, *Briefings . Bioinf.* (2016). <http://dx.doi.org/10.1093/bib/bbw067>.
- [77] B.T. McInnes, T. Pedersen, Evaluating measures of semantic similarity and relatedness to disambiguate terms in biomedical text, *J. Biomed. Inf.* 46 (6) (2013) 1116–1124, doi:10.1016/j.jbi.2013.08.008.
- [78] B.T. McInnes, T. Pedersen, S.V.S. Pakhomov, UMLS-interface and UMLS-similarity: open source software for measuring paths and semantic similarity, in: *Proceedings of the Annual Symposium of the American Medical Informatics Association*, 2009, ncbi.nlm.nih.gov, San Francisco, CA, 2009, pp. 431–435.
- [79] K. Mehlhorn, P. Sanders, *Algorithms and Data Structures: The Basic Toolbox*, SpringerLink: Springer e-Books, Springer, 2008.
- [80] J. Mendling, H.A. Reijers, J. Recker, Activity labeling in process modeling: empirical insights and recommendations, *Inf. Syst.* 35 (4) (2010) 467–482, doi:10.1016/j.is.2009.03.009.
- [81] L. Meng, J. Gu, A new model for measuring word sense similarity in WordNet, in: *Proceedings of the 4th International Conference on Advanced Communication and Networking, ASTL*, 14, 2012, pp. 18–23.
- [82] L. Meng, J. Gu, Z. Zhou, A new model of information content based on concept's topology for measuring semantic similarity in WordNet, *Int. J. Grid Distrib. Comput.* 5 (3) (2012) 81–93.
- [83] L. Meng, R. Huang, J. Gu, Measuring semantic similarity of word pairs using path and information content, *Int. J. Future Gener. Commun. Netw.* 7 (3) (2014) 183–194, doi:10.14257/ijfgcn.2014.7.3.17.
- [84] D. Merkel, Docker: lightweight linux containers for consistent development and deployment, *Linux J.* 2014 (239) (2014), Article No. 2.
- [85] R. Meymandpour, J.G. Davis, A semantic similarity measure for linked data: an information content-based approach, *Knowl.-Based Syst.* 109 (2016) 276–293, doi:10.1016/j.knsys.2016.07.012.
- [86] R. Mihalcea, C. Corley, C. Strapparava, Corpus-based and knowledge-based measures of text semantic similarity, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, 1, AAAI Press, 2006, pp. 775–780.
- [87] G.A. Miller, WordNet: a lexical database for English, *Commun. ACM* 38 (11) (1995) 39–41.
- [88] G.A. Miller, W.G. Charles, Contextual correlates of semantic similarity, *Lang. Cognit. Processes* 6 (1) (1991) 1–28, doi:10.1080/01690969108406936.
- [89] S. Montani, G. Leonardi, Retrieval and clustering for supporting business process adjustment and analysis, *Inf. Syst.* 40 (2014) 128–141, doi:10.1016/j.is.2012.11.006.
- [90] M.R. Munafò, B.A. Nosek, D.V.M. Bishop, K.S. Button, C.D. Chambers, N.P. du Sert, U. Simonsohn, E.-J. Wagenmakers, J.J. Ware, J.P.A. Ioannidis, A manifesto for reproducible science, *Nat. Hum. Behav.* 1 (2017) 0021, doi:10.1038/s41562-016-0021.
- [91] J. Oliva, J.I. Serrano, M.D. del Castillo, A. Iglesias, SyMSS: A syntax-based measure for short-text semantic similarity, *Data Knowl. Eng.* 70 (4) (2011) 390–405, doi:10.1016/j.datak.2011.01.002.
- [92] S. Patwardhan, S. Banerjee, T. Pedersen, Using measures of semantic relatedness for word sense disambiguation, in: A. Gelbukh (Ed.), *Proceedings of the 4th International Conference on Computational Linguistics and Intelligent Text Processing (CICLING 2003)*, LNCS, 2588, Springer, Mexico D.F., 2003, pp. 241–257, doi:10.1007/3-540-36456-0_24.
- [93] S. Patwardhan, T. Pedersen, Using WordNet-based context vectors to estimate the semantic relatedness of concepts, in: *Proceedings of the EAACL 2006 Workshop Making Sense of Sense-Bringing Computational Linguistics and Psycholinguistics Together*, 1501, 2006, pp. 1–8. Trento, Italy.
- [94] T. Pedersen, Empiricism is not a matter of faith, *Comput. Ling.* 34 (3) (2008) 465–470, doi:10.1162/coli.2008.34.3.465.
- [95] T. Pedersen, in: *WordNet-InfoContent-3.0.tar dataset repository*, 2008. (https://www.researchgate.net/publication/273885902_WordNet-InfoContent-3.0.tar).
- [96] T. Pedersen, Information Content Measures of Semantic Similarity Perform Better Without Sense-tagged Text, in: *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, in: HLT '10, Association for Computational Linguistics, Stroudsburg, PA, USA, 2010, pp. 329–332.
- [97] T. Pedersen, in: Measuring the similarity and relatedness of concepts: a MICAI 2013 Tutorial, 2013, doi:10.13140/RG.2.1.3025.6164.
- [98] T. Pedersen, S.V.S. Pakhomov, S. Patwardhan, C.G. Chute, Measures of semantic similarity and relatedness in the biomedical domain, *J. Biomed. Inf.* 40 (3) (2007) 288–299, doi:10.1016/j.jbi.2006.06.004.
- [99] T. Pedersen, S. Patwardhan, J. Michelizzi, WordNet::similarity: measuring the relatedness of concepts, in: *Demonstration Papers at HLT-NAACL 2004*, in: HLT-NAACL-Demonstrations '04, Association for Computational Linguistics, Stroudsburg, PA, USA, 2004, pp. 38–41.

- [100] V. Pekar, S. Staab, Taxonomy learning: factoring the structure of a taxonomy into a semantic classification decision, in: Proceedings of the 19th International Conference on Computational Linguistics, in: COLING '02, 1, Association for Computational Linguistics, Stroudsburg, PA, USA, 2002, pp. 1–7, doi:[10.3115/1072228.1072318](https://doi.org/10.3115/1072228.1072318).
- [101] C. Pesquita, D. Faria, A.O. Falcao, P. Lord, F.M. Couto, Semantic similarity in biomedical ontologies, *PLoS Comput. Biol.* 5 (7) (2009) e1000443, doi:[10.1371/journal.pcbi.1000443](https://doi.org/10.1371/journal.pcbi.1000443).
- [102] E. Petrakis, G. Varelas, A. Hliaoutakis, P. Raftopoulou, X-similarity: computing semantic similarity between concepts from different ontologies, *J. Digital Inf. Manage.* 4 (4) (2006) 233–237.
- [103] G. Pirró, A semantic similarity metric combining features and intrinsic information content, *Data . Knowl. Eng.* 68 (11) (2009) 1289–1308, doi:[10.1016/j.datak.2009.06.008](https://doi.org/10.1016/j.datak.2009.06.008).
- [104] G. Pirró, J. Euzenat, A feature and information theoretic framework for semantic similarity and relatedness, in: P.F. Patel-Schneider, Y. Pan, P. Hitzler, P. Mika, L. Zhang, J.Z. Pan, I. Horrocks, B. Glimm (Eds.), Proceedings of the 9th International Semantic Web Conference, ISWC 2010, LNCS, 6496, Springer, Shanghai, China, 2010, pp. 615–630, doi:[10.1007/978-3-642-17746-0_39](https://doi.org/10.1007/978-3-642-17746-0_39).
- [105] G. Pirró, N. Seco, Design, implementation and evaluation of a new semantic similarity metric combining features and intrinsic information content, in: R. Meersman, Z. Tari (Eds.), On the Move to Meaningful Internet Systems: OTM 2008, LNCS, 5332, Springer, 2008, pp. 1271–1288, doi:[10.1007/978-3-540-88873-4_25](https://doi.org/10.1007/978-3-540-88873-4_25).
- [106] E.M. Pothos, J.R. Busemeyer, J.S. Trueblood, A quantum geometric model of similarity, *Psychol. Rev.* 120 (3) (2013) 679–696, doi:[10.1037/a0033142](https://doi.org/10.1037/a0033142).
- [107] R. Rada, H. Mili, E. Bicknell, M. Blettner, Development and application of a metric on semantic nets, *IEEE Trans. Syst. Man. Cybern.* 19 (1) (1989) 17–30, doi:[10.1109/21.24528](https://doi.org/10.1109/21.24528).
- [108] P. Resnik, Using information content to evaluate semantic similarity in a taxonomy, in: Proceedings of the International Joint Conferences on Artificial Intelligence (IJCAI 1995), 1, 1995, pp. 448–453, Montreal, Canada.
- [109] P. Resnik, Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language, *J. Artif. Intell. Res.* 11 (1999) 95–130.
- [110] M.A. Rodríguez, M.J. Egenhofer, Determining semantic similarity among entity classes from different ontologies, *IEEE Trans. Knowl. Data Eng.* 15 (2) (2003) 442–456, doi:[10.1109/TKDE.2003.1185844](https://doi.org/10.1109/TKDE.2003.1185844).
- [111] H. Rubenstein, J.B. Goodenough, Contextual correlates of synonymy, *Commun. ACM* 8 (10) (1965) 627–633, doi:[10.1145/365628.365657](https://doi.org/10.1145/365628.365657).
- [112] D. Sánchez, M. Batet, Semantic similarity estimation in the biomedical domain: an ontology-based information-theoretic perspective, *J. Biomed. Inf.* 44 (5) (2011) 749–759, doi:[10.1016/j.jbi.2011.03.013](https://doi.org/10.1016/j.jbi.2011.03.013).
- [113] D. Sánchez, M. Batet, A new model to compute the information content of concepts from taxonomic knowledge, *Int. J. Seman. Web Inf. Syst. (ISWIS)* 8 (2) (2012) 34–50, doi:[10.4018/jswis.2012040102](https://doi.org/10.4018/jswis.2012040102).
- [114] D. Sánchez, M. Batet, D. Isern, Ontology-based information content computation, *Knowl.-Based Syst.* 24 (2) (2011) 297–303, doi:[10.1016/j.knsys.2010.10.001](https://doi.org/10.1016/j.knsys.2010.10.001).
- [115] D. Sánchez, M. Batet, D. Isern, A. Valls, Ontology-based semantic similarity: a new feature-based approach, *Expert Syst. Appl.* 39 (9) (2012) 7718–7728, doi:[10.1016/j.eswa.2012.01.082](https://doi.org/10.1016/j.eswa.2012.01.082).
- [116] A. Schlicker, F.S. Domingues, J. Rahnenführer, T. Lengauer, A new measure for functional similarity of gene products based on Gene Ontology, *BMC Bioinf.* 7 (2006) 302, doi:[10.1186/1471-2105-7-302](https://doi.org/10.1186/1471-2105-7-302).
- [117] A. Schlicker, T. Lengauer, M. Albrecht, Improving disease gene prioritization using the semantic similarity of Gene Ontology terms, *Bioinformatics* 26 (18) (2010). i561–7 [10.1093/bioinformatics/btq384](https://doi.org/10.1093/bioinformatics/btq384)
- [118] A. Sebt, A.A. Barfroush, A new word sense similarity measure in WordNet, in: Proceedings of the International Multiconference on Computer Science and Information Technology, IMCSIT 2008, IEEE, 2008, pp. 369–373, doi:[10.1109/IMCSIT.2008.4747267](https://doi.org/10.1109/IMCSIT.2008.4747267).
- [119] N. Seco, T. Veale, J. Hayes, An intrinsic information content metric for semantic similarity in WordNet, in: R. López de Mántaras, L. Saitta (Eds.), Proceedings of the 16th European Conference on Artificial Intelligence (ECAI), 16, IOS Press, Valencia, Spain, 2004, pp. 1089–1094.
- [120] M.H. Seddiqui, M. Aono, Metric of intrinsic information content for measuring semantic similarity in an ontology, in: Proceedings of the Seventh Asia-Pacific Conference on Conceptual Modelling, in: APCCM '10, 110, Australian Computer Society, Inc., Darlinghurst, Australia, 2010, pp. 89–96.
- [121] H. Shima, in: WS4J home page, 2011. (<https://code.google.com/p/ws4j/>).
- [122] L. Stanchev, Creating a similarity graph from WordNet, in: Proceedings of the 4th International Conference on Web Intelligence, Mining and Semantics (WIMS'14). Article No. 36, ACM, 2014, doi:[10.1145/2611040.2611055](https://doi.org/10.1145/2611040.2611055).
- [123] N. Stojanovic, A. Maedche, S. Staab, R. Studer, Y. Sure, SEAL: a framework for developing sEMantic portALS, in: Proceedings of the 1st International Conference on Knowledge Capture, in: K-CAP '01, ACM, New York, NY, USA, 2001, pp. 155–162, doi:[10.1145/500737.500762](https://doi.org/10.1145/500737.500762).
- [124] A. Tversky, Features of similarity, *Psychol. Rev.* 84 (4) (1977) 327–352, doi:[10.1037/0033-295X.84.4.327](https://doi.org/10.1037/0033-295X.84.4.327).
- [125] E. Van Miltenburg, WordNet-based similarity metrics for adjectives, in: Proceedings of the 8th Global WordNet Conference, Global WordNet Association, Bucharest, Romania, 2016, pp. 414–418.
- [126] D. Vrandečić, M. Krötzsch, Wikidata: a free collaborative knowledge base, *Commun. ACM* 57 (10) (2014) 78–85, doi:[10.1145/2629489](https://doi.org/10.1145/2629489).
- [127] A. Wolke, M. Bichler, F. Chirigati, V. Steeves, Reproducible experiments on dynamic resource allocation in cloud data centers, *Inf. Syst.* 59 (2016) 98–101, doi:[10.1016/j.is.2015.12.004](https://doi.org/10.1016/j.is.2015.12.004).
- [128] A. Wolke, B. Tsend-Ayush, C. Pfeiffer, M. Bichler, More than bin packing: dynamic resource allocation strategies in cloud data centers, *Inf. Syst.* 52 (2015) 83–95, doi:[10.1016/j.is.2015.03.003](https://doi.org/10.1016/j.is.2015.03.003).
- [129] X. Wu, E. Pang, K. Lin, Z.-M. Pei, Improving the measurement of semantic similarity between gene ontology terms and gene products: insights from an edge- and IC-based hybrid method, *PLoS One* 8 (5) (2013) e66745, doi:[10.1371/journal.pone.0066745](https://doi.org/10.1371/journal.pone.0066745).
- [130] Z. Wu, M. Palmer, Verbs semantics and lexical selection, in: Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics, in: ACL '94, Association for Computational Linguistics, Stroudsburg, PA, USA, 1994, pp. 133–138, doi:[10.3115/981732.981751](https://doi.org/10.3115/981732.981751).
- [131] Q. Yuan, Z. Yu, K. Wang, A new model of information content for measuring the semantic similarity between concepts, in: Proceedings of the International Conference on Cloud Computing and Big Data (CloudCom-Asia 2013), IEEE Computer Society, 2013, pp. 141–146, doi:[10.1109/CLOUDCOM-ASIA.2013.25](https://doi.org/10.1109/CLOUDCOM-ASIA.2013.25).
- [132] S.-B. Zhang, J.-H. Lai, Exploring information from the topology beneath the Gene Ontology terms to improve semantic similarity measures, *Gene* 586 (1) (2016) 148–157, doi:[10.1016/j.gene.2016.04.024](https://doi.org/10.1016/j.gene.2016.04.024).
- [133] Z. Zhou, Y. Wang, J. Gu, A new model of information content for semantic similarity in WordNet, in: Proceedings of the Second International Conference on Future Generation Communication and Networking Symposia (FGCNS'08), 3, IEEE, 2008, pp. 85–89, doi:[10.1109/FGCNS.2008.16](https://doi.org/10.1109/FGCNS.2008.16).
- [134] Z. Zhou, Y. Wang, J. Gu, New model of semantic similarity measuring in WordNet, in: Proceedings of the 3rd International Conference on Intelligent System and Knowledge Engineering (ISKE 2008), 1, IEEE, 2008, pp. 256–261, doi:[10.1109/ISKE.2008.4730937](https://doi.org/10.1109/ISKE.2008.4730937).

Chapter 11

US Patent Application

This page intentionally left blank.



US 20160179945A1

(19) **United States**

(12) **Patent Application Publication**

LASTRA DIAZ et al.

(10) **Pub. No.: US 2016/0179945 A1**

(43) **Pub. Date: Jun. 23, 2016**

(54) **SYSTEM AND METHOD FOR THE INDEXING AND RETRIEVAL OF SEMANTICALLY ANNOTATED DATA USING AN ONTOLOGY-BASED INFORMATION RETRIEVAL MODEL**

(52) **U.S. Cl.**
CPC *G06F 17/30734* (2013.01); *G06F 17/30595* (2013.01); *G06F 17/3043* (2013.01); *G06F 17/30525* (2013.01); *G06F 17/3053* (2013.01); *G06F 17/2785* (2013.01)

(71) Applicant: **UNIVERSIDAD NACIONAL DE EDUCACIÓN A DISTANCIA (UNED)**, Madrid (ES)

(72) Inventors: **Juan Jose LASTRA DIAZ**, Madrid (ES); **Ana GARCIA SERRANO**, Madrid (ES)

(21) Appl. No.: **14/576,679**

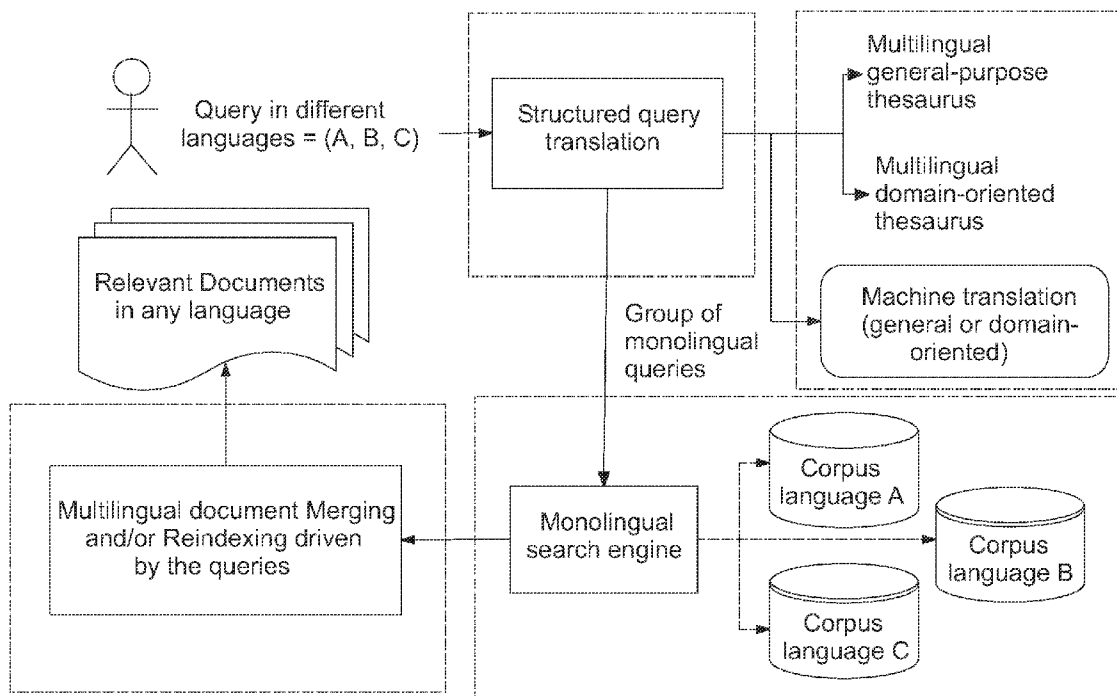
(22) Filed: **Dec. 19, 2014**

Publication Classification

(51) **Int. Cl.**
G06F 17/30 (2006.01)
G06F 17/27 (2006.01)

(57) **ABSTRACT**

System and method for the indexing and retrieval of semantically annotated information units from a collection of semantically annotated indexed information units in response to a query using an ontology-based IR model. The retrieval method comprises: receiving a semantically annotated query with semantic annotations to individuals or classes within a determined populated base ontology; embedding, as a set of weighted-mentions to individuals or classes within the populated base ontology, the semantically annotated query in a semantic representation space of an ontology-based metric space IR model; obtaining the representation in the semantic representation space for every indexed information unit; computing the Hausdorff distance between the space representation of the query and the space representation of all the indexed information units of the collection; retrieving and ranking, the relevant information units based on the computed Hausdorff distance.



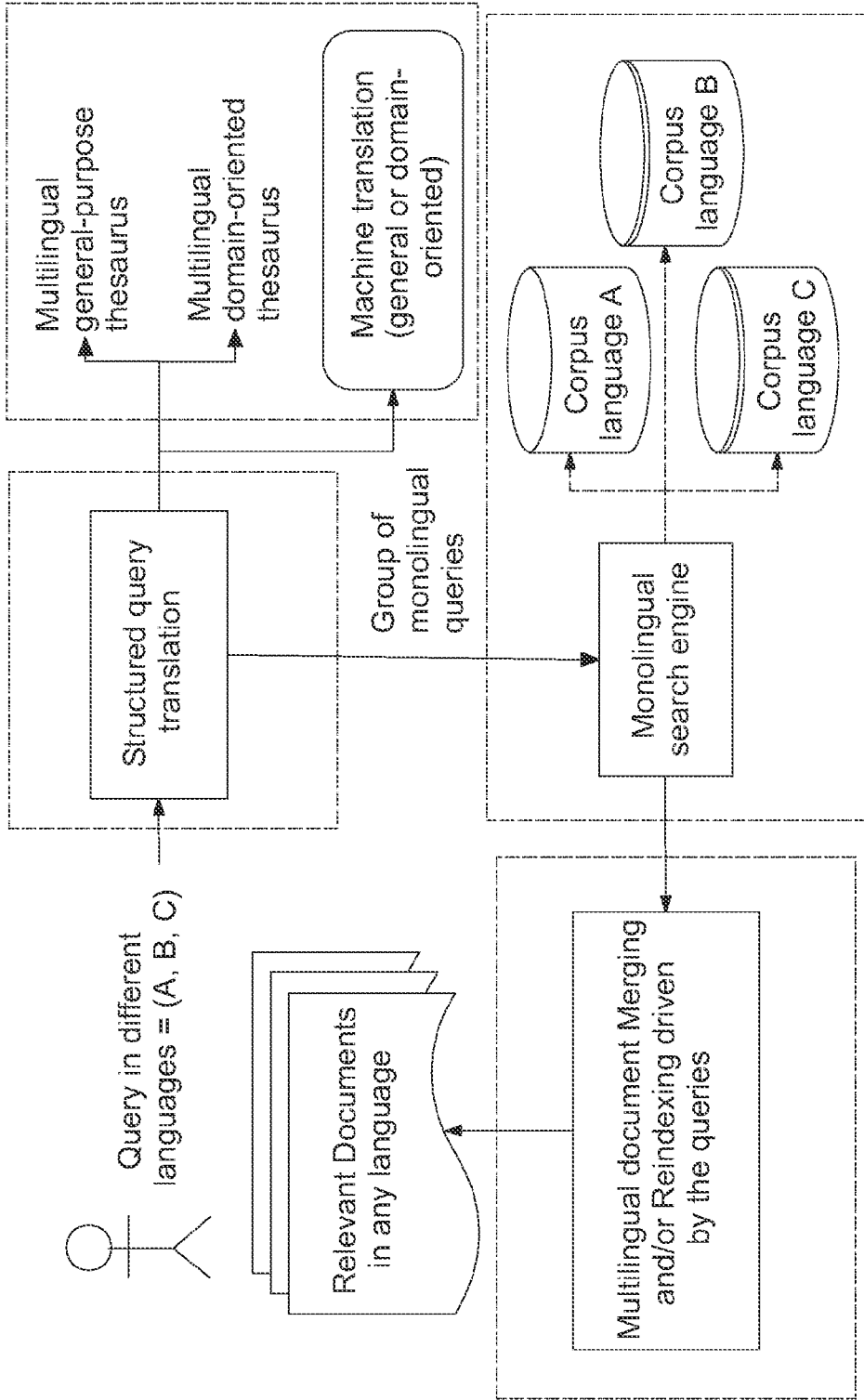


Fig. 1

Ontology-based metric space models					
IR Model	Doc. features	Doc. Space	Retrieval	Weighting	Ranking
Rada et al., 1989	Set of boolean concepts	Ontology-based metric space (shortest path)	Integrated in ranking	Boolean	Average distance among sets of concepts
Present invention	Set of weighed instances and concepts	Intrinsic ontology-based metric space (extension of JC distance)	Integrated in ranking	IC-based TF weights	Hausdorff distance among sets of weighted concepts and instances
Concept-based adapted VSM models					
IR Model	Doc. features	Doc. Space	Retrieval	Weighting	Ranking
Fang et al., 2005	Weighted bivector of instances and keywords	Instance-keyword based bivector VSM	OWL-DL queries	TFIDF + semantic saliency factor	Bivector cosine score combination
Vallet et al., 2005 Castells et al., 2007	Weighted bivector of instances and keywords	Instance-keyword based bivector VSM	SPARQL queries	TFIDF	Bivector cosine score combination
Mustafa et al., 2008	Concept-based weighted vector	Concept-based single vector VSM	Pairwise semantic distance among sets of concepts (query and document)	TFIDF	Combined score (cosine function + edge-counting semantic distance)
Dragoni et al., 2010	Wordnet concept-based vector	Concept-based single vector VSM	Integrated in ranking	TFIDF	Cosine score
Egozi et al., 2011	Vector of weighed keywords - Wikipedia ESA-concepts	enriched ESA concept-based single vector VSM	ESA-based retrieval (doc-passages ranking)	Combined ESA-concept cosine score for passages and full document	Combined concept-based cosine score document-passages vs query
Cao & Ngo, 2012	Multi-vectors of weighted keywords + ontological features	enriched concept-based multi-vector VSM	Integrated in ranking	TFIDF	Barycentric combination of multiple cosine scores
Machhour & Kassou, 2013	Vector of weighted concepts	concept-based single vector VSM	Integrated in ranking	TFIDF	Cosine score

Fig. 2

Measure	Semantic similarity or distance measure
Rada et al.	$d(c_1, c_2) = \min_{all\ paths} \{L(c_1, c_2)\}$
Wu-Palmer	$sim(c_1, c_2) = \frac{2ds(c_1, v_2)}{ds(c_1) + ds(c_2)}$
Hirst-St-Onge	$d(c_1, c_2) = \frac{L(c_1, c_2)}{k}$
Leacock-Chodorow	$d(c_1, c_2) = \frac{L(c_1, c_2)}{\max_{e_i \in C} \{ds(e_i)\}}$
Resnik	$sim(c_1, c_2) = \max_{e_i \in \text{sup}(c_1, c_2)} \{IC(e_i)\}$
Jiang-Conrath	$d(c_1, c_2) = IC(c_1) + IC(c_2) - 2IC(LCA(c_1, c_2))$
Lin	$sim(c_1, c_2) = \frac{2 \ln(\psi(LCA(c_1, c_2)))}{\ln(\psi(c_1)) + \ln(\psi(c_2))}$
Distance in the present invention	$d_{w,JC}(c_1, c_2) = \min_{z \in Paths(c_1, c_2)} \left\{ \sum_{e_{ij} \in z} w(e_{ij}) \right\}$ $w(e_{ij}) = IC(P(v_i v_j))$

Fig. 3

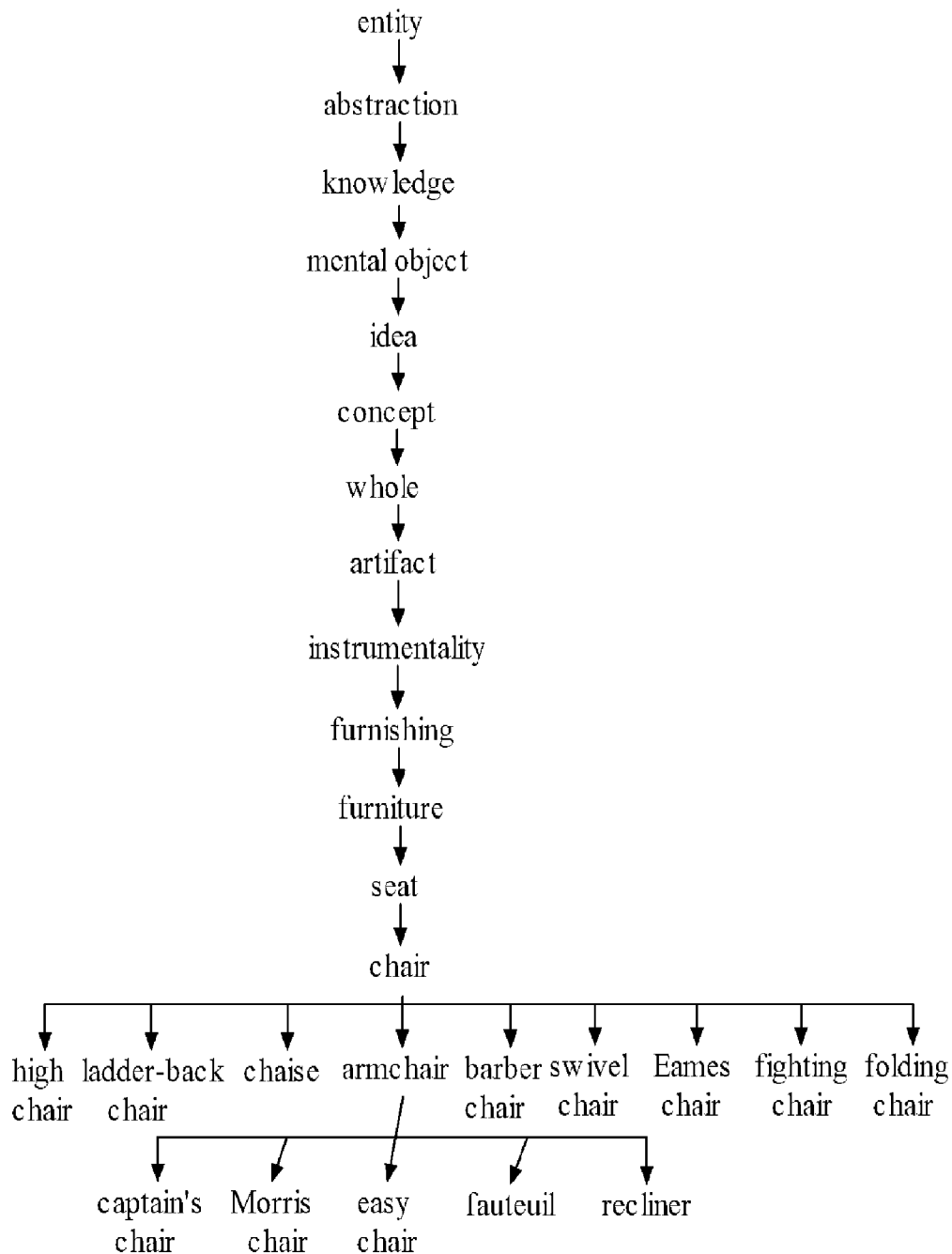


Fig. 4

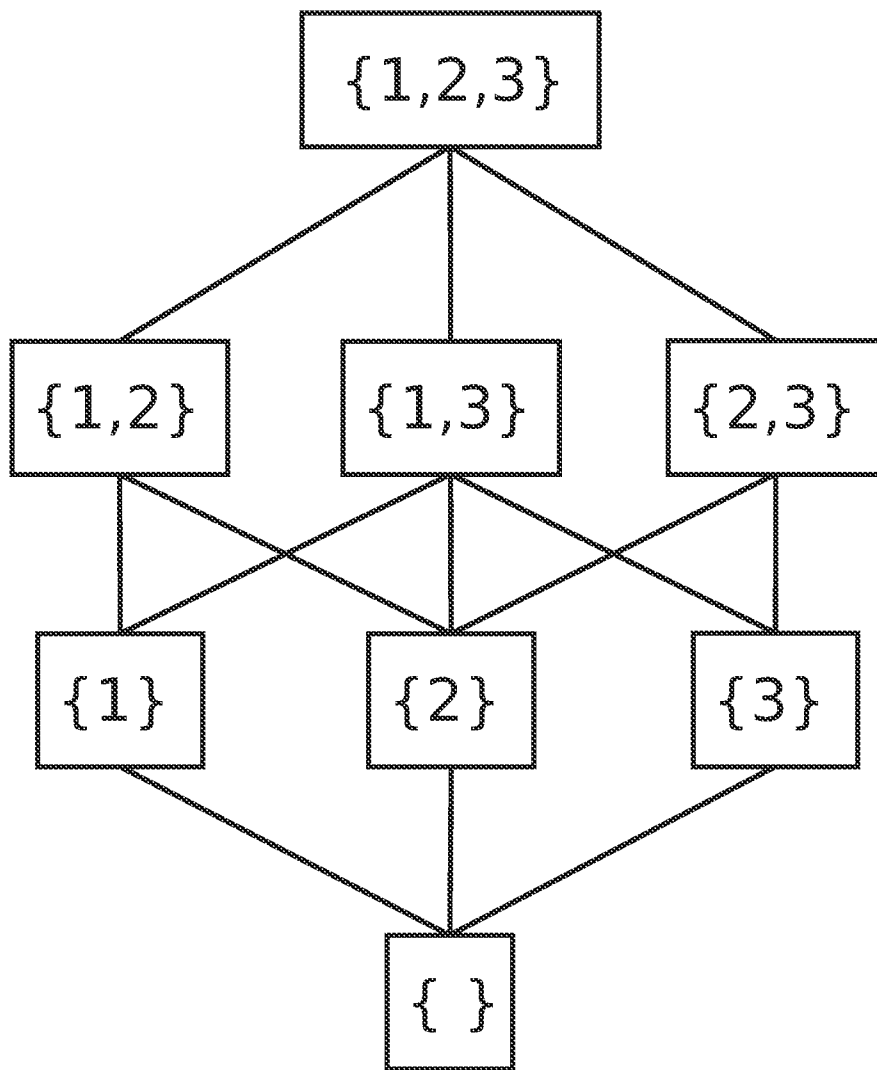


Fig. 5

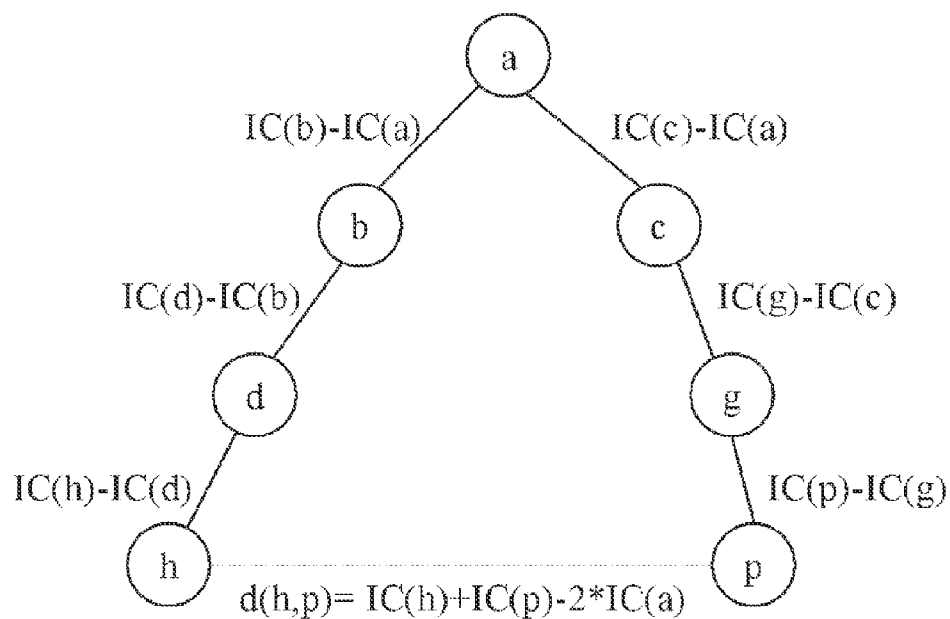


Fig. 6

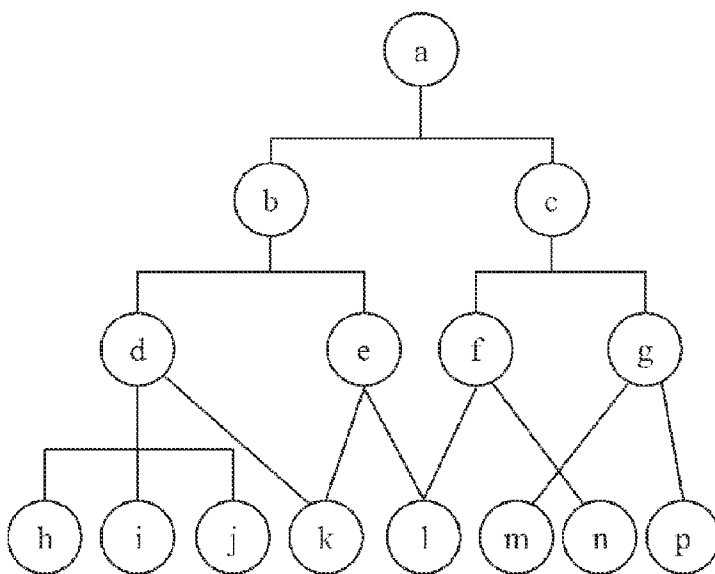


Fig. 7

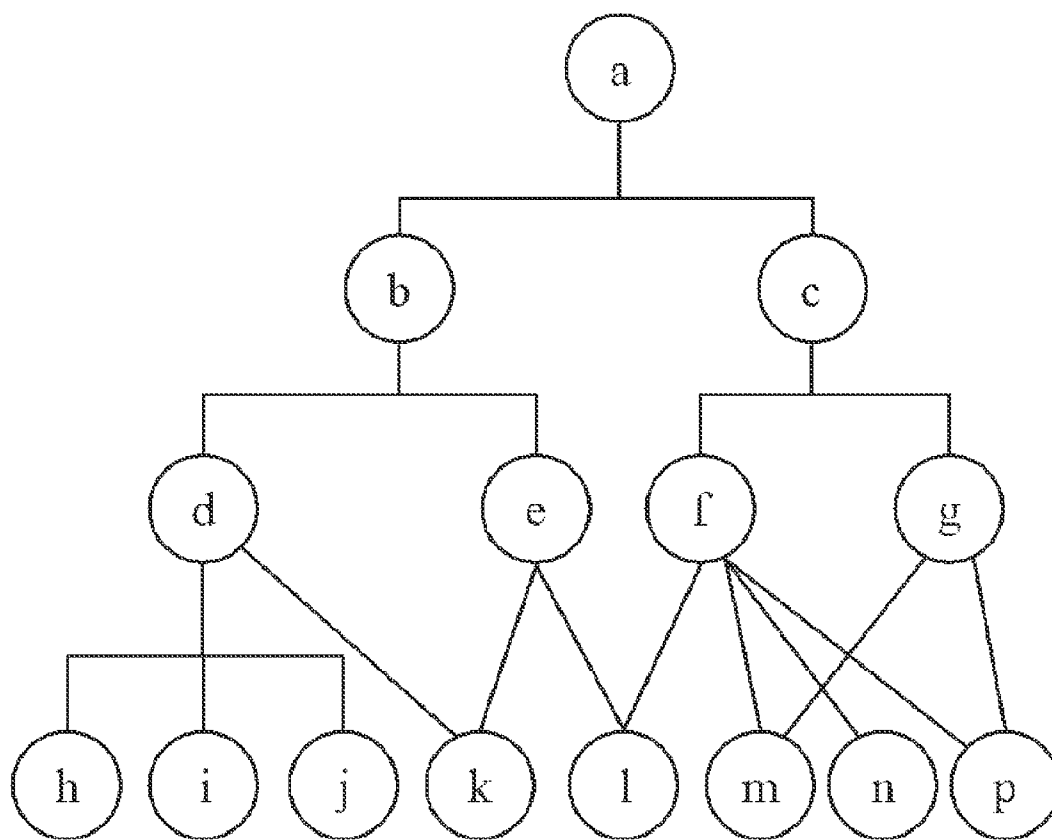


Fig. 8

Id	Elements of the IOS model	Definition and notation
1	Input base ontology (taxonomy)	$T = (C, \leq_C)$ with a root element $\rho \in C \mid \forall c \in C \rightarrow c \leq_C \rho$
2	Family of sets of instances of any ontology class	$I_C = \{I_C\}$
3	Populated base ontology	$O = (C \cup I_C, \leq_C)$
4	Metric ontology	$O = (C \cup I_C, \leq_C, d_C)$
5	Frequency-based weighted-set space	$D = C \cup I_C \times \mathbb{N}$
6	D of semantic annotations	$\delta_k = \{(\tau_j, f_j^k) \in C \cup I_C \times \mathbb{N} \mid j \in J(k)\}$
7	Input information unit $\delta_k \subset D$ defined by a tuple of frequency-based semantic annotations within the base ontology.	(X, d_X) , with $X = C \cup I_C \times [0, 1] \subset \mathbb{R}$ and d_X is a metric
8	Ontological representation space given by the metric space (X, d_X)	A pair of functions $\Phi = (\varphi_I, \varphi_C)$, that verifies the next axioms: (1) $C_1 \leq_C C_2 \Rightarrow \varphi_C(C_1) \subset \varphi_C(C_2), \forall (C_1, C_2) \in C \times C$ (2) $d_C(C_1, C_2) = d_H(\varphi_C(C_1), \varphi_C(C_2)), \forall (C_1, C_2) \in C \times C$ (3) $\varphi_I(\tau) \subset \varphi_C(C_i), \forall \tau \in I_C,$ $\varphi_I(\tau) \subset \varphi_C(C_j), \forall \tau \in I_C, \forall C_j \mid C_i \leq_C C_j$
9	Intrinsic ontology embedding (structure-preserving)	$\varphi_I : D \rightarrow X$ $\varphi_I(\tau_j, f_j^k) = \begin{cases} (\tau_j, 1), & \forall \tau_j \in I_C \\ (e_{C_i}, 1) & \forall \tau_j \in C \end{cases}$
10	Intrinsic embedding φ_I for individuals in the proposed model	$\varphi_C : C \rightarrow X$ $\varphi_C(C_i) = \{x \in X \mid \pi_O(x) \leq_C C_i\}$ $\pi_O : X \rightarrow C \cup I_C$ $\pi_O(\tau_j, w_j) = \tau_j$
	Class embedding φ_C for mention to full classes (query) in the proposed model.	

Fig. 9A

Id	Elements of the IOS model	Definition and notation
11	Information units embedding φ and static weights w_j^k (indexes)	Frequency-based annotations space: $D = C \cup I_C \times \mathbb{N}$ Freq-based indexed units space (power set on D): $\mathcal{D} = \mathcal{P}(D)$ $\varphi: \mathcal{D} \rightarrow X$ $\varphi(\delta_k) = \{(\tau_j, \omega_j^k) \in X \mid j \in J(k)\}, \quad \omega_j^k = \frac{f_j^k}{\sum_{j \in J(k)} f_j^k}$
12	Semantic weighting (normalized IC value for indexed information units) which endorses our TF-based weighting scheme in row 11 above.	$\widehat{IC}: \mathcal{D} \rightarrow \mathbb{R}$ $\widehat{IC}(\delta_k) = \sum_{j \in J(k)} \underbrace{\omega_j^k}_{\text{static weight}} \cdot \underbrace{IC(\tau_j)}_{\text{dynamic \& semantic}}$
13	Novel weighted Jiang-Comraish distance among concepts, denoted by d_{wJC} . The values are pre-computed in the pre-processing step (see Figure 11).	$d_{wJC}: \mathcal{O} \times \mathcal{O} \rightarrow \mathbb{R}$ $d_{wJC}(a, b) = \min_{z \in \mathcal{P}(a, b)} \left\{ \sum_{e_{ij} \in z} w(e_{ij}) \right\}$ $w: E \rightarrow \mathbb{R}$ $w(e_{ij}) = IC(\mathcal{P}(c_i c_j)) = -\log_2(\mathcal{P}(c_i c_j))$
14	Metric of the IOS representation space (distance among weighted-mentions to individuals and classes). This function is used in block (6.1) of Figure 12.	$d_X: X \times X \rightarrow \mathbb{R}$ $d_X((x, w_x), (y, w_y)) = \begin{cases} \min_{LA(x) \setminus LA(y)} \left\{ \begin{array}{l} \underbrace{-\log_2(w_x \cdot \mathcal{P}(x LA(x)))}_{(1)} \\ \underbrace{-\log_2(w_y \cdot \mathcal{P}(y LA(y)))}_{(2)} \\ \underbrace{+d_{wJC}(LA(x), LA(y))}_{(3)} \end{array} \right\}, & \text{if } x \neq y \\ \left \log_2 \left(\frac{w_x}{w_y} \right) \right , & \text{if } x = y \end{cases}$
15	Intrinsic Ontological Spaces (IOS) = metric space + intrinsic embedding	(X, d_X) , with d_X defined in (14) above
16	Hausdorff distance among subsets of a metric space (X, d_X)	$d_H: \mathcal{P}((X, d_X)) \times \mathcal{P}((X, d_X))$ $d_H(A, B) = \max \left\{ \sup_{a \in A} \{d_X(a, B)\}, \sup_{b \in B} \{d_X(b, A)\} \right\}$
17	Ranking of the model given by the est-valued distance function d_D . This function is used in block (6) of Figure 12.	$d_D: \mathcal{D} \times \mathcal{D} \rightarrow \mathbb{R}$ $d_D(\delta_k, \delta_m) = d_H(\varphi(\delta_k), \varphi(\delta_m))$
18	Distance in the representation space between any weighted-mention and the image of a full class (query mention), denoted by d_{IC} . It is a special case of d_X that reduces to the expression here. This function is used in block (6.2) of Figure 12.	$d_{IC}: X \times \varphi_C \subset X \rightarrow \mathbb{R}$ $d_{IC}((\tau_j, \omega_j^k), \varphi_C(C_i)) = \begin{cases} 0, & \tau_j \leq C_i \\ d_X((\tau_j, \omega_j), (x_n, w_j)), & \text{otherwise} \end{cases}$

Fig. 9B

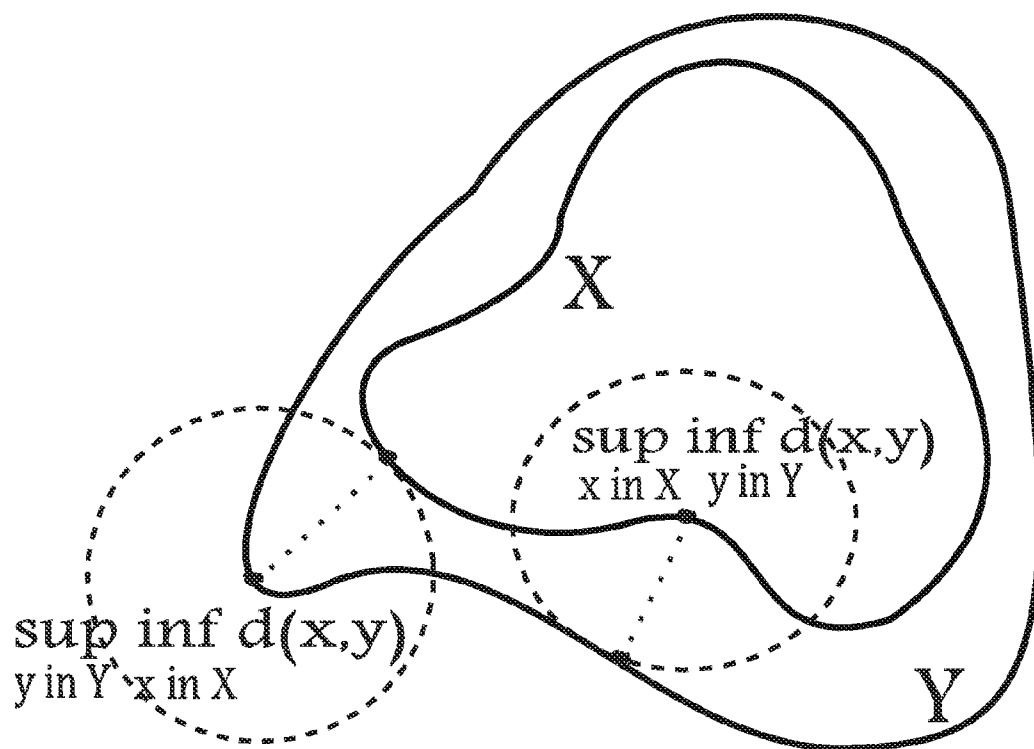


Fig. 10

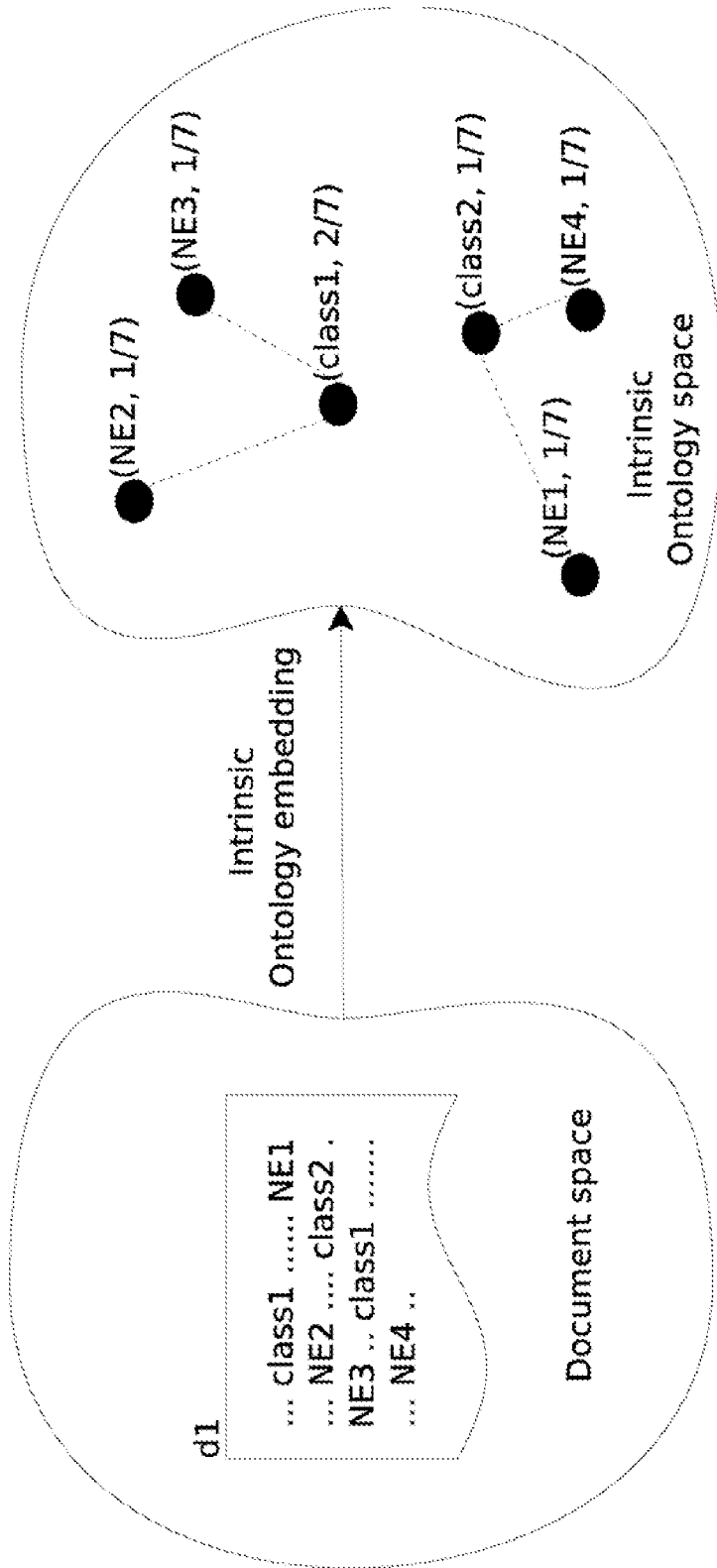


Fig. 11

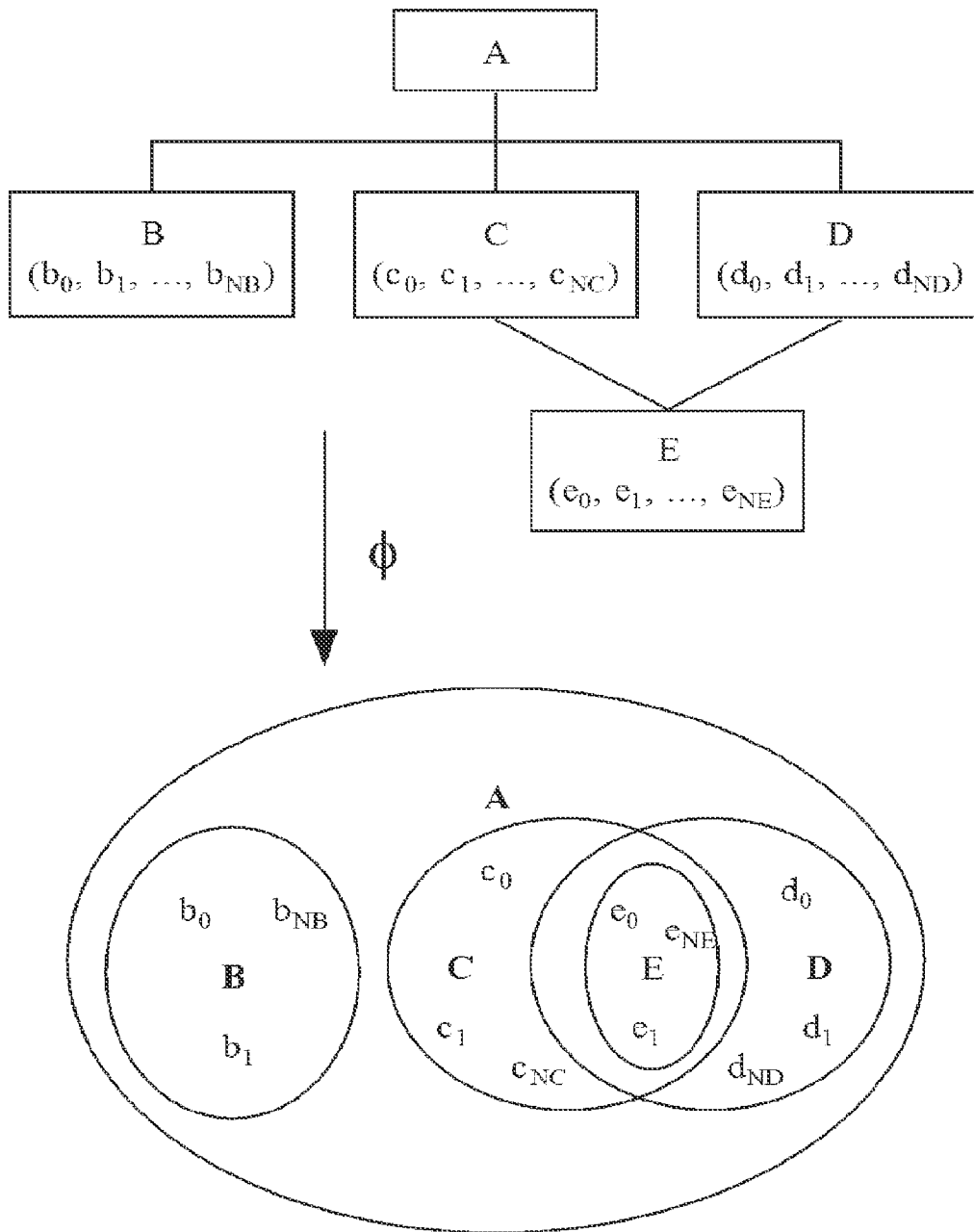


Fig. 12

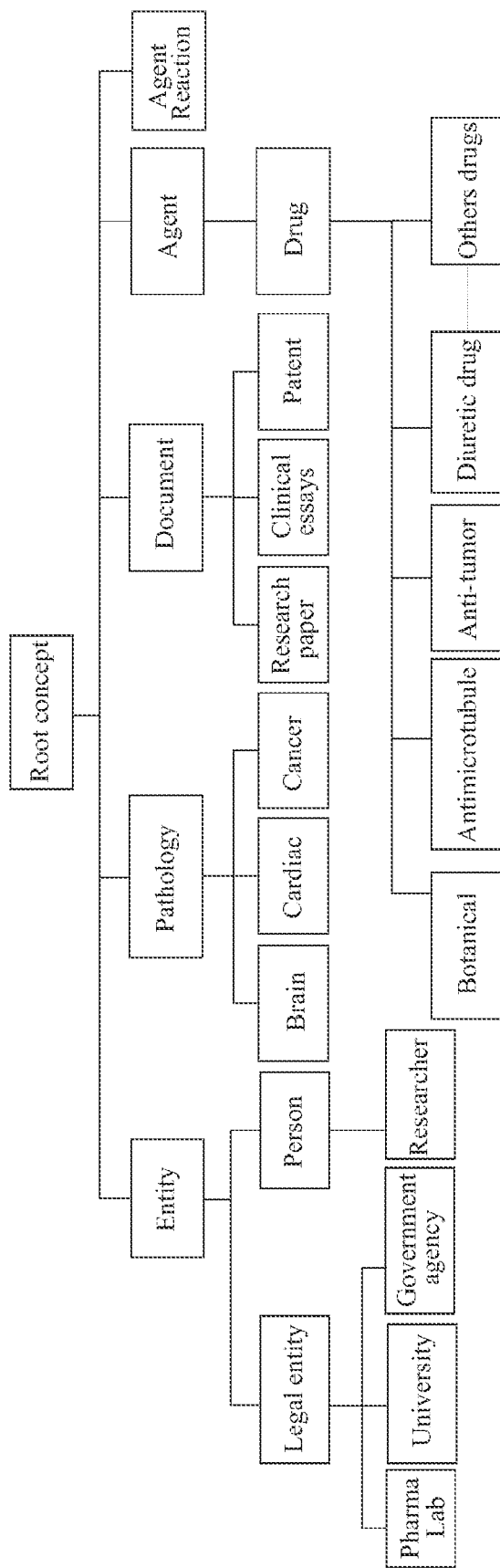


Fig. 13

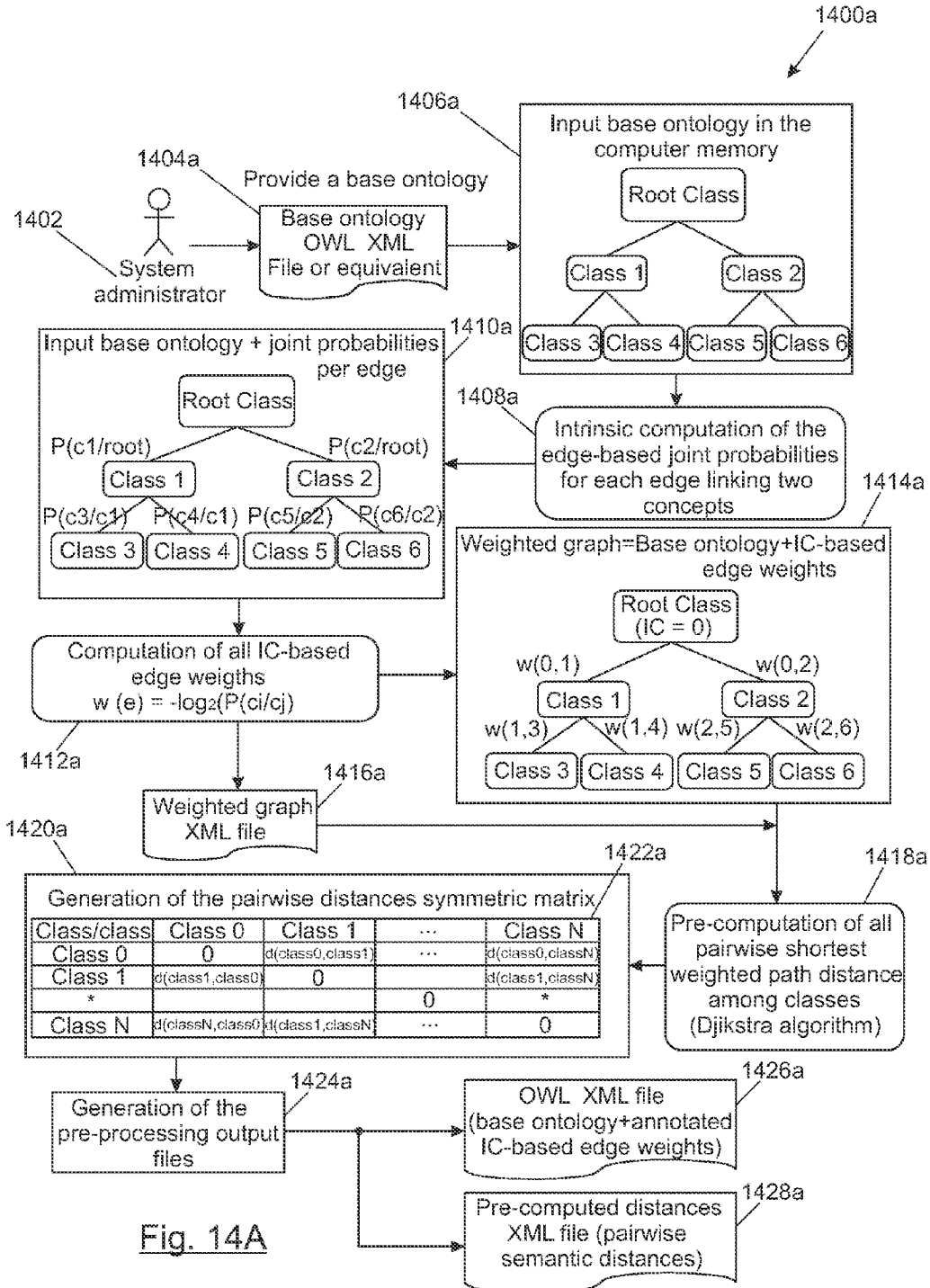


Fig. 14A

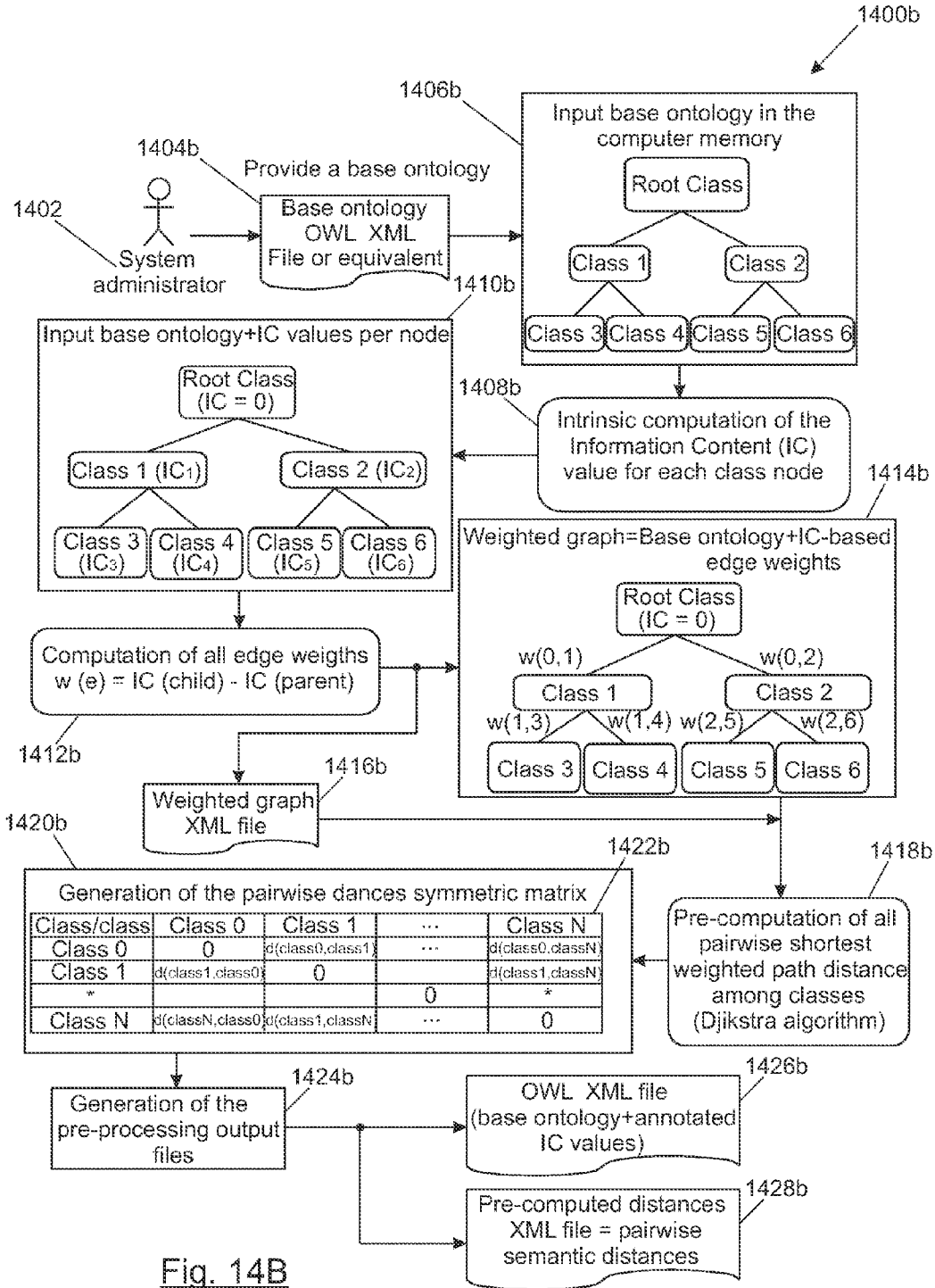


Fig. 14B

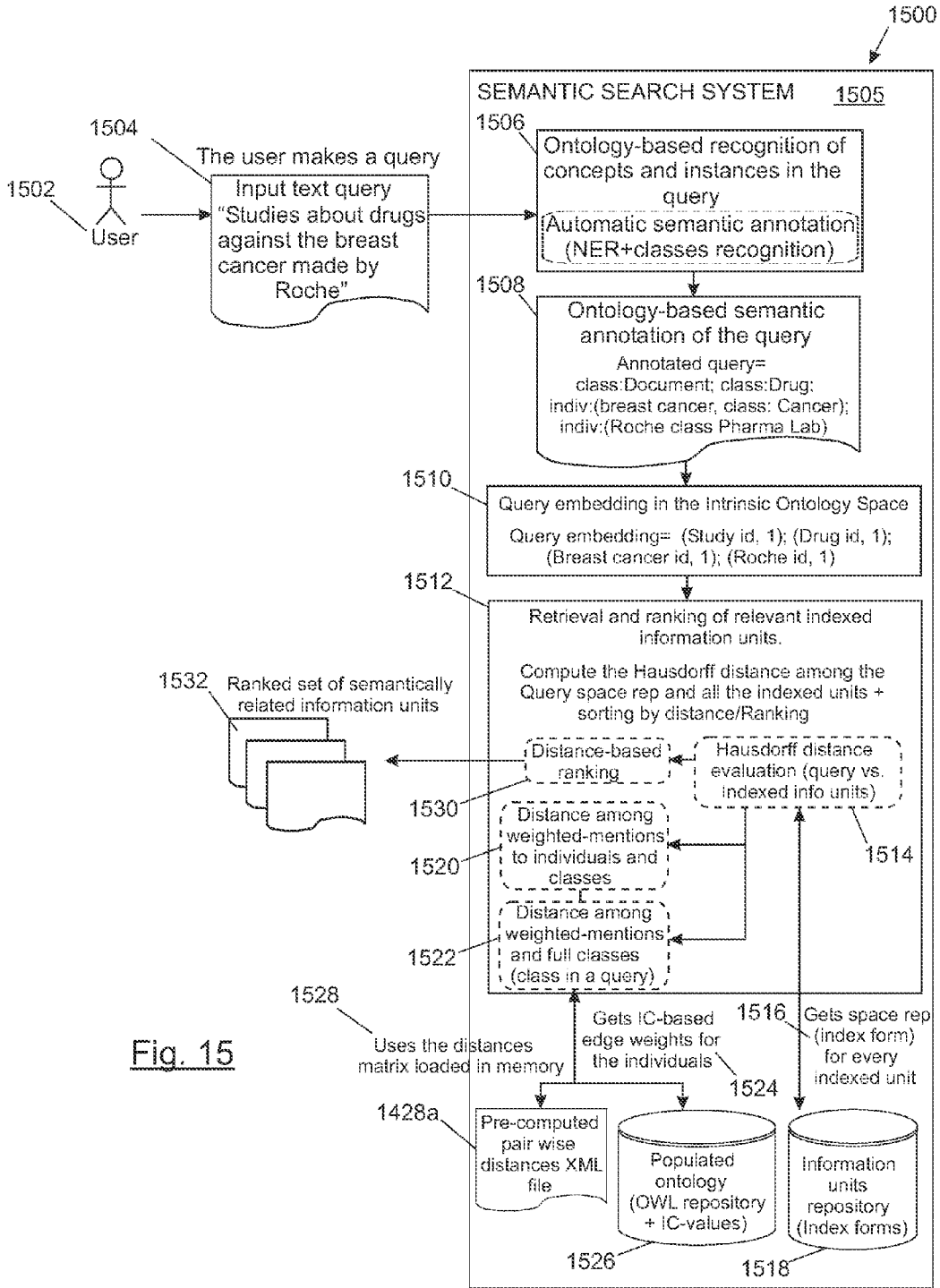
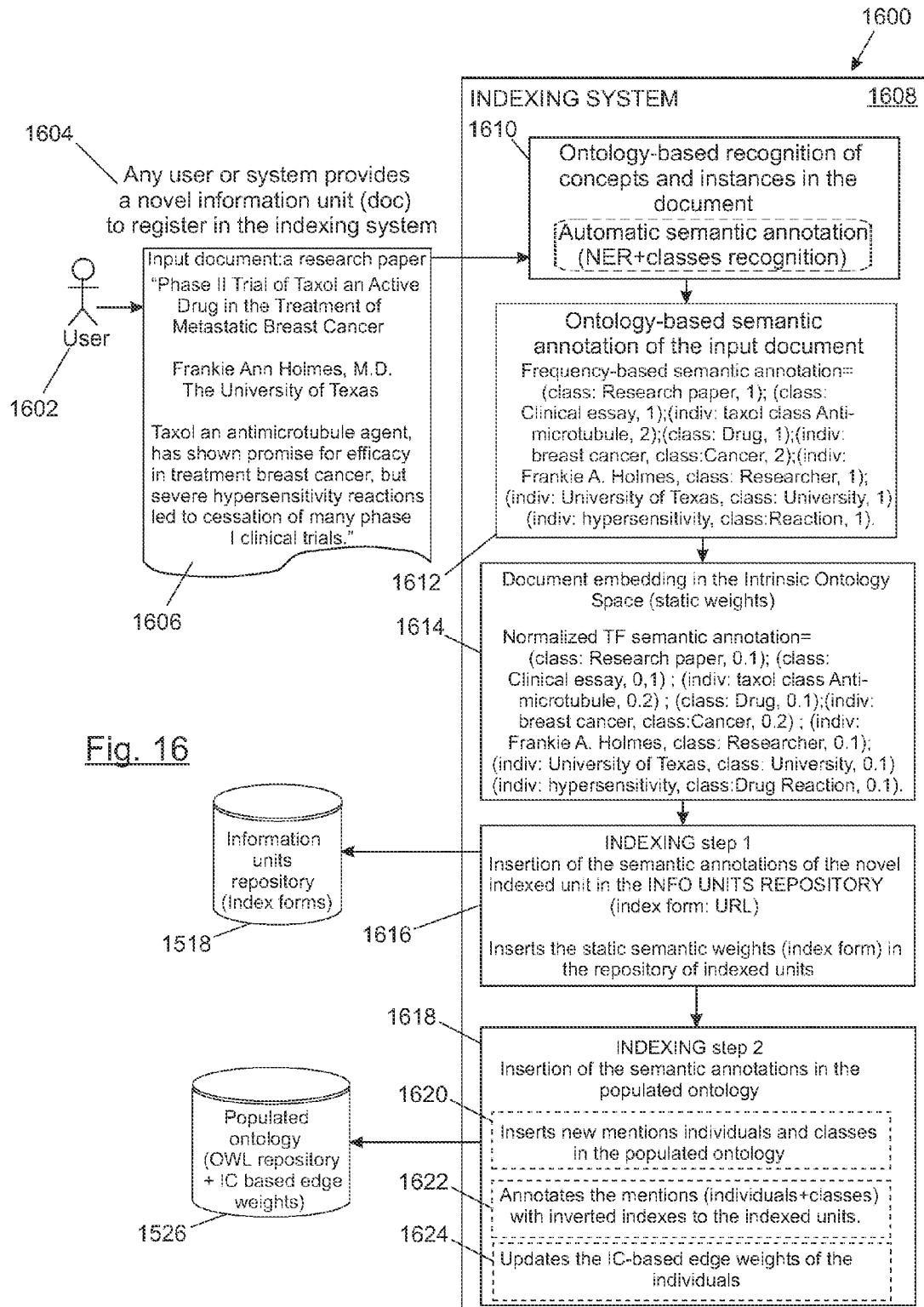


Fig. 15



**SYSTEM AND METHOD FOR THE INDEXING
AND RETRIEVAL OF SEMANTICALLY
ANNOTATED DATA USING AN
ONTOLOGY-BASED INFORMATION
RETRIEVAL MODEL**

FIELD OF THE INVENTION

[0001] The present invention pertains to the field of information retrieval (IR). It particularly relates to the disclosure of a novel method and system to build a structure-preserving ontology-based IR model, called Intrinsic Ontological Spaces, for the indexing and retrieval of semantically annotated data, such as text documents, web pages, or any sort of information that can be represented as a set of semantic annotations (individuals and classes) on any sort of base ontology. The proposed method bridges the gap of modelling inconsistencies in current methods through the integration of the intrinsic structure of any populated ontology in the definition of the representation space itself. Our approach can be interpreted as a semantic metrization of the populated ontologies used to represent the semantic annotations of the indexed data. Moreover, the present invention also discloses a novel ontology-based semantic distance called weighted Jiang-Conrath distance, and a family of three novel intrinsic Information Content (IC) methods for the computation of the IC-based edge weights used in the weighted Jiang-Conrath distance. These three novel intrinsic IC-computation methods are herewith called IC-JointProbUniform, IC-JointProbHypo and IC-JointProbLeaves.

[0002] The broader use of the method proposed in the invention is the development of semantic search systems for any sort of information sources that have been semantically annotated within a base ontology. The main benefit for the final users, or the public, is that any search system based in the proposed invention will get an improvement in terms of the ranking quality of the retrieved information units, and the precision and recall measures. As result, the joint improvement of these evaluation measures contributes to improve the results expected by the final users of any information search system.

BACKGROUND OF THE INVENTION

[0003] The object of the present invention is a novel ontology-based IR indexing and retrieval method and system for any sort of semantically annotated data, called in this invention as Intrinsic Ontological Spaces. The semantically annotated data are also called information units to emphasize that the model can be used to index any sort of data, such as text documents, web pages, images or any other sort of multimedia data. One of the most broadly available semantically annotated information units are the collections of text documents, or web pages, and the main application of the proposed model will be the indexation, retrieval and ranking of text documents relevant for any semantically annotated query.

[0004] The proposed method is framed in the family of ontology-based Information Retrieval (IR) models, and its main motivation is to overcome the drawbacks of the current ontology-based IR models reported in the literature.

[0005] The Vector Space Model (VSM) [Salton et al., 1975] is known as “bag of words”, because every document is represented by a vector whose coordinates are defined as a function of the term occurrence frequency within a document. The set of terms used to represent every document is called

the vocabulary of the model, and it defines the base vectors of the model. In most of cases, the cosine function is used as a similarity measure between a query vector and the vectors representing the indexed documents. Due to its simplicity, the VSM model has been adopted in many tasks and applications of natural language processing (NLP), such as: information retrieval (IR), document categorization (TC) and clustering, web mining and automatic text summarization (TS) among others.

[0006] Recently, the vector space models has been extended to define word and phrase spaces, such is reflected in reviews made in [Erk, 2012], [Clark, 2012] and [Turney & Pantel, 2010]. A word or phrase space is a vector space where the vectors represent these information units instead of documents, and the space metric encodes the semantic similarity between information unit pairs. The word spaces are based in the distributional hypothesis [Basin & Pennacchiotti, 2010], which sets that words in similar contexts have similar meanings. In these models, the vectors representing every word are built as a function of the terms frequency in the context of one word within a document, so that these models allow encoding some semantic relations and statistics, such as the term co-occurrence, the synonymy and the meronymy among others.

[0007] Although the vector space models has been mainly used to represent text documents, such as we saw above, these models have been successfully applied to represent other types of information units, such as words, phrases and sentences. Following the previous reasoning, the ontology-based IR model proposed here works with any information unit that can be encoded in an ontology, according to the definition below.

[0008] The information units are the objects indexed by the ontology-based IR model proposed in the present invention, and it could be text documents, web pages, sentences, multimedia objects, or any sort of data that admit a concept-instances ontology representation. In the context of the present invention, an information unit is defined as any sort of semantically annotated data that can be represented as a collection of concepts (classes) or instances of them (individuals) within an ontology.

[0009] The main limitation of the VSM model is its lack of meaning. As it is noted in [Metzler, 2007, pp. 3], most of the current academic information retrieval models use a standard “bag of words” VSM model with meaningless terms, which make impossible to retrieve documents using queries with non-explicitly terms mentioned in the corpus. By other hand, the same situation occurs in other related problems where the same meaningless version of the VSM model was used, by example, in the text categorization problem [Sebastiani, 2002] [Lewis et al., 2004]. As it is noted in [Metzler, 2007], little information is known about the IR models used in commercial search engines like Yahoo, or Google, however, given the results to some input query, we could think that these models are mainly based in meaningless terms.

[0010] The advent of the semantic web has motivated a great change of paradigm in the IR community, the IR models has moved from a model based in meaningless terms to a model based on references to concepts or its instances. This new paradigm has converted the conceptual models and the knowledge bases in its core components, and ontology languages, such as OWL, has become the favorite representation to encode this knowledge and to store the references to the indexed data. Nowadays, the use of ontologies is omnipresent in all kind of semantic retrieval task in the context of the

semantic web [Ding et al., 2007], as well as in other application contexts as the bioinformatics [Pesquita et al., 2009].

[0011] Motivated by the lack of meaning in previous IR models, some novel conceptual IR models have appeared during the last decade, whose main example is the family of ontology-based IR models. The abstract definition of these IR models is given below.

[0012] An ontology-based IR model is any sort of information retrieval model which uses an ontology-based conceptual representation for the content of any sort of information unit, whose main goal is its indexing, retrieval and ranking regarding to a user query.

[0013] The family of ontology-based IR models can be subdivided in three subfamilies:

[0014] (1) the vector ontology-based IR models, such as those disclosed in [Vallet et al., 2005], [Fang et al., 2005], [Castells et al., 2007], [Mustafa et al., 2008], [Dragoni et al., 2010] and [Egozi et al., 2011], whose main feature is the use of some adaptation of the standard VSM model to manage concepts instead of meaningless terms,

[0015] (2) the ontology-based metric space IR models, whose unique known examples are the pioneering work of [Rada et al., 1989] and the present invention, and

[0016] (3) the query-expansion ontology-based IR models, such as those disclosed in U.S. Pat. No. 8,301,633 B2 and U.S. Pat. No. 6,675,159 B1.

[0017] The only known ontology-based IR model based in a metric space is the model proposed in [Rada et al., 1989]. The work in [Rada et al., 1989] can be considered as the oldest reference within the ontology-based IR family. However, this work is not cited by the ontology-based IR models found in the literature, despite that the Rada's measure is highly cited and well known in the literature about ontology-based semantic distances.

[0018] The main features of the family of vector ontology-based IR models, also called adapted-VSM models, are:

[0019] (1) the use of a conceptual representation for documents and queries based in an ontology,

[0020] (2) the retrieval of relevant documents through any ontology query language,

[0021] (3) some sort of vector space for the representation of references to concepts and instances, based in a set of orthogonal base vectors defined by the classes and individuals of the ontology,

[0022] (4) some sort of adaptation of standard term-frequency weights for the definition of coordinates,

[0023] (5) the use of cosine function as ranking method to sort the relevant documents, and

[0024] (6) a multivector representation and ranking combining different types of features, such as concepts, keywords or ontological features.

[0025] A vector space is a very rich algebraic structure that, precisely by its richness, has been underused or misused in the scope of information retrieval. Formally, a vector space is an additive Abelian group with a scalar product that is associative and distributive, it means that the space vector includes all the inverse elements for each document, and every linear combination between them; nevertheless, all these elements of the space are not used, or required, in any IR model. Actually, the only reason to use vector spaces in the current IR models is to rank the documents using the cosine function as similarity measure, due to its simplicity and computational efficiency.

[0026] The state of the art in ontology-based IR models has proven the potential benefits derived from the use of conceptual models with regard to the meaningless IR models. However, if we study carefully the assumptions made by these conceptual models, we find some important aspects that offer an important improvement capability in terms of ranking quality, as well as in the precision and recall measures.

[0027] Main motivation behind most of the adapted-VSM models have been to build a semantic weighting method to compare semantically annotated documents, however, these models have been using the vector space model (VSM) as a black-box without take into account some important implicit assumptions of the model and its consequences.

[0028] Making a review of the current literature about the topic, we find the following gap which motivates the present invention. The gap is summarized in seven issues:

[0029] (1) orthogonality condition,

[0030] (2) cardinality mismatch,

[0031] (3) statistical fingerprint vs. semantic distances,

[0032] (4) populated ontologies are not directly indexed,

[0033] (5) lack of a semantic weighting,

[0034] (6) continuity problems of some proposed metrics on sets, (7) the Jiang-Conrath distance is not a well defined metric, and

[0035] (8) the Jiang-Conrath distance cannot be directly applied to sets of weighted-mentions to classes and individuals.

[0036] (9) the current intrinsic IC-computation methods, used in combination with any ontology-based IC-based semantic measure as the disclosed here, does not fulfil the following structural constraints: the difference of the IC values between a parent concept and a child concept in a taxonomy must be equal to the joint probability between them, and the sum of the joint probabilities of the children concepts in every parent concept must be equal to 1.

[0037] Orthogonality condition. The base vectors of any VSM model are mutually orthogonal, it means that similarity cosine function between different base vectors is zero. One consequence of the orthogonality condition of the adapted VSM models is that two vectors associated with two documents can get a zero, or very low similarity value, when they do not share references to the same concept instances, although these instances could share a common ancestor concept in the taxonomy. For example, documents with references to bicycle and motorbike models would not be related, although the instances are derived from the two-wheel vehicle concept.

[0038] Cardinality mismatch. Most of these ontology-based models are not including references to classes as sets of objects, and others are mixing references to classes and instances (individuals) at the same representation level. The main idea behind most adaptations of the VSM models to manage the ontology information is to make a mapping from individuals and/or classes to base vectors of the representation vector space. In this way, the models are assigning two different and opposite meanings to the same base vector, in one case the base vector represents the occurrence of one object (individuals), while in the opposite case, a base vector is representing a collection of objects (classes). These inconsistencies can be summarized as a cardinality mismatch in the adapted VSM models, and the nature of the objects represented by the model.

[0039] Statistical fingerprint vs. semantic distances. The metric used to compare documents by most of published

ontology-based models is based in the Euclidean angle between normalized vectors (cosine score). The vectors encode the statistical fingerprint of the indexed documents (i.e. the statistical co-occurrence relations between different concepts in a document), but this metric lacks of a meaning in the sense that they are not encoding any semantic distance between concepts, as it is made by very well established ontology-based distances, such as the Jiang-Conrath distance measure [Jiang & Conrath, 1997]. The only exception to this problem is the IR model proposed in [Rada et al., 1989] which defines a Boolean semantic model, where the documents are represented by sets of concepts, but the concepts are annotated in binary form without using any semantic weighting method, as it is provided by the method of the present invention.

[0040] Populated ontologies are not directly indexed. Many of the ontology-based IR VSM models need to retrieve the related documents with the instances and concepts in the query before ranking them. The populated ontology is not indexed directly; for this reason, it needs to be searched using any ontology-based query language, such as SPARQL or any other. By contrast, the model of the present invention builds a direct geometric representation of the data in the populated ontology, integrating retrieval and ranking in a same step. It can produce bottlenecks for large scale ontologies, but the geometric model allows the integration of well-established geometrical search structures to speed-up the queries, such as the introduced in [Brin et al., 1995].

[0041] Lack of a semantic weighting. The weights in adapted VSM models are statistical values, not related to the real semantic weight of the concept/instance in the document.

[0042] Continuity problems of some proposed metrics on sets. In [Rada et al., 1989], the authors introduce an ontology-based IR model which defines a metric space using a shortest path metric on the taxonomy, and the average distance between sets of concepts as a distance between documents. The authors report some continuity problems which can be attributed to the use of a not well defined metric on sets.

[0043] The Jiang-Conrath distance is not a well-defined metric. Some recent research has unveiled that the Jiang-Conrath distance only satisfies the metric axioms for tree-like ontologies [Orum & Joslyn, 2009]. This fact contradicts the original statement of the authors in [Jiang & Conrath, 1997]. The Jiang-Conrath distance depends on the lowest common ancestor between two concepts, which is only uniquely defined for lattices, not for general partially ordered sets (posets). Despite of the JC distance is well defined on lattices, in [Orum & Joslyn, 2009] the authors provide some counter-examples to demonstrate that neither in this case the JC distance is a metric. The novel ontology-based semantic distance disclosed in the present invention introduces a generalization of the Jiang-Conrath distance to fulfil the metric axioms on any sort of taxonomy, solving the drawbacks described above.

[0044] The Jiang-Conrath cannot be directly applied to sets of weighted-mentions to classes and individuals, as required by the semantic representation space introduced herein. The Intrinsic Ontological Spaces of the present invention defines a metric space which unifies the representation of weighted classes and individuals in a same space; thus, the model needs a semantic distance that can be extended from concepts in a taxonomy to weighted classes and individuals associate to the same objects into a populated ontology. The novel weighted Jiang-Conrath disclosed herein allows to bridge this gap,

providing a base metric to define the metric of the whole representation space, which combines four classes of elements:

- [0045]** (1) weighted-mentions to classes,
- [0046]** (2) weighted-mentions to individuals,
- [0047]** (3) whole-mentions to classes as sets, and
- [0048]** (4) whole-mentions to individuals.

[0049] The current intrinsic IC-computation methods, such as the introduced in [Seco et al., 2004] and [Zhou et al., 2008] do not fulfil some important constraints related to the definition of the taxonomy as the domain of a probability space. In [Jiang & Conrath, 1997], the authors note the relation between the difference of IC values between any two adjacent concepts {child,parent} in a taxonomy T, and the joint probability $P(\text{child}|\text{parent})$. Precisely, this relation has been taken into account before in the literature to design new semantic measures, and we use it to derive our novel semantic distance. Specifically, this relation is given by $IC(\text{child})-IC(\text{parent})=-\text{binaryLog}(P(\text{child}|\text{parent}))$. Other important probabilistic constraint is that the sum of the $P(\text{child}|\text{parent})$ values for each parent node must be equal to 1. To bridge this gap we have designed one family of intrinsic IC-computation methods based in the intrinsic estimation of the joint probability that we disclose in this invention, and we use to compute the IC-based edge weights required by our novel ontology-based semantic distance. Although these IC-computation methods are used here with our novel edge-based IC semantic distance, they can be directly adapted to use in combination with any ontology-based node-based IC semantic measures, such as the measures introduced in [Resnik, 1995], [Jiang & Conrath, 1997] or [Lin, 1998] among others. Therefore, these IC-computation methods have other direct application moreover the ontology-based IR model disclosed in this invention.

[0050] The use of any sort of vector space models is omnipresent in all sorts of information retrieval (IR) models for all sorts of web and data search engines. The ontology-based IR model proposed in this invention defines a new paradigm for the semantic indexing of all sort of semantically annotated data, whose main goal is transforming the search processes made by the users from a keyword-based search to a concept-based search. Therefore, this invention can be considered as a complement and a new generation of IR models destined to substitute the current generation of keyword-based search engines. The method disclosed in this invention is framed in the family of ontology-based IR models, and it shares a common goal with other previous methods: be the cornerstone of a new generation of semantic search systems. As the VSM models, the proposed invention can be applied in the context of any natural language processing (NLP) application where any sort of semantic space is used, among we can cite: web search system, any sort of IR system for text indexing and retrieval, cross-language information retrieval (CLIR) systems, automatic text summarization systems, text categorization and clustering, question answering systems, and word disambiguation among others. Moreover, the proposed model also can be applied in bioengineering applications where the data and the domain knowledge are represented within a domain-oriented ontology.

[0051] The ontology-based IR model proposed in the present invention is able to update any sort of application based in semantic vector spaces, or ontology-based adapted VSM models. By example, the VSM model has been extensively used in the context of multilingual or cross-language IR systems (MLIR/CLIR), such as the model shown in FIG. 1.

Most of CLIR systems are based in classical monolingual IR models, which are interrogated using translations of the input queries. The monolingual retrieved documents are re-indexed and/or merged to be finally ranked according to its saliency for the input query.

[0052] Other problem where adaptations of the VSM model have been proven its utility, and the proposed model in the present invention could be applied, is in the automatic text summarization (TS). In the scope of extractive TS methods we can find some conceptual models based in adaptations of the VSM model to represent the semantic similarity relations among sentences. This is the followed approach in [Meng et al., 2005] where the authors define the conceptual vector space model (CVSM), whose ideas are very close to the ontology-based IR model approach. Other TS methods are based in the clustering of sentences, originating the notion of centrality, whose core idea is that any document can be represented by the more significative (central) sentence. These clustering methods use a VSM model to represent the sentences in any document, where each vector encodes a set of features of the sentence, and the model can use different functions to establish the similarity among sentences. Among these clustering TS methods we find the pioneering works of [McKeown et al., 1999] and [Hatzivassiloglou et al., 2001], as well as the works in [Siddharthan et al., 2004] and [García-Hernández & Ledeneva, 2009]. Finally, the most recent text summarization (TS) methods are based in graph-ranking algorithms derived from PageRank and HITS, whose main references are the works of [Erkan & Radev, 2004], [Mihalcea & Tarau, 2004], [Wolf & Ginson, 2004] and [Vanderwende et al., 2004]. If the sentences within a document are considered as information units, these graph-based methods could benefit from the proposed model in the present invention, because the graphs are derived from the semantic similarity among sentences obtained through adaptations of a vector space model and a set of semantic features.

[0053] In the scope of the Q&A systems, the vector space models have been used to represent sentences within a document, and to retrieve text fragments with a potential answer to a question. These approaches inspired in IR models, jointly with other techniques, have been successfully proven in DeepQA [Chu-Carroll et al., 2012] to retrieve relevant text fragments for a user query.

[0054] Finally, other potential application of the Intrinsic Ontological Spaces of the present invention is the word disambiguation problem, where several methods based in the vector representation of the context of a word have been proposed [Navigli, 2009], following the distributional hypothesis. Due to the omnipresence of the vector space models in NLP, the proposed model has many potential applications in the scope of NLP applications.

[0055] Some related work is hereinafter described in the context of this invention as well as some potentially related patents.

[0056] The present invention is mainly related with three categories of works in the literature:

[0057] (1) the ontology-based IR models,

[0058] (2) geometric representations for taxonomies, and

[0059] (3) ontology-based semantic distances.

[0060] According to the research problem studied, the present invention pertains to the family of ontology-based IR models, whereas according to the approach followed in the proposed solution, the present invention is strongly related with the family of ontology-based semantic distance and

similarity measures. In fact, the present invention includes a novel ontology-based semantic distance called weighted Jiang-Conrath distance. By other hand, the geometric approach adopted in the present invention is inspired by the geometric spirit in the pioneering works about geometry and meaning of [Widdows, 2004] and [Clarke, 2007].

[0061] For a better understanding of the literature about the topic, a summary of the main features of the analysed ontology-based IR models are shown in FIG. 2. The ontology-based IR models have been categorized in three subfamilies according to the structure of its representation space: (1) metric-space models, like in the present invention, (2) adapted Vectors Space Models (VSM-based), and (3) query-expansion models, represented by some patents cited herein. In FIG. 2, we only include a comparison of the first two subfamilies of ontology-based IR models, while the works in family of query-expansion models are subsequently described in detail.

[0062] Hereinafter, the state of the art about the ontology-based IR models will first be reviewed. Later, some methods and ideas for the geometric representation of taxonomies that are related to the core ideas of the IR model proposed in the present invention will also be commented. Lastly, the state of the art about ontology-based semantic distances will be introduced and reviewed, enumerating the known facts and drawbacks about the Jiang-Conrath distance, which have motivated the development of the novel semantic distance that we call weighted Jiang-Conrath distance.

1. Ontology-Based IR Models

[0063] Regarding the ontology-based IR models in the state of the art, some previous surveys about the family of ontology-based IR models can be found, such as the reviews made in [Castells, 2008] and [Fernandez et al., 2011], as well as the survey in the context of multimedia retrieval made in [Kannan et al., 2012]. In other work [Wu et al., 2011], the authors survey the query expansion problem in IR and others ontology-based IR models, which will be later discussed. The surveys cited can be useful to follow the analysis of the state of the art about the object of the invention herein disclosed.

[0064] Ontology-Based IR Models Versus the Query Expansion Approaches.

[0065] The query expansion problem was reviewed in [Wu et al., 2011]. The relation between this approach and the ontology-based IR models are analysed. We can consider the query expansion approach the dual of the ontology-based models: the first one expands the query, while the last one expands the conceptual representation of the document. In the query expansion approach, the terms in the query are expanded with synonyms, related concepts or semantic annotations, and the expanded vector of terms is used to interrogate an unstructured semantic representation space. By other hand, in the ontology-based approach, the representation space is structured, and the semantic relations are already implicit in the indexation model, therefore the semantic representation of the documents is already expanded to match the queries in its base form.

[0066] In [Castells, 2008], the author makes a literature survey about the use of ontologies in IR and web mining, approach commonly known as semantic web, while he describes his experience in the development of a ontology-based IR system introduced in [Castells et al., 2007]. In other recent work [Fernandez et al., 2011], the same group of authors introduce some extensions to the model in [Castells et

al., 2007] to can operate at web scale, while they also extend their previous literature survey. In [Kannan et al., 2012], the authors survey the ontology-based IR models in the context of the multimedia IR field.

[0067] Ontology-Based IR Models Based in Metric Spaces.

[0068] The first published ontology-based IR model is proposed in [Rada et al., 1989]. The main motivation of this work is the development of an IR model for biomedical applications, where the documents are represented as sets of concepts within a common ontology. Rada et al. proposes to use the shortest path between concepts within an ontology as a measure of its semantic distance, and they call this measure “distance”. The proposed IR model represents the documents and the queries by the set of concepts referenced in these information sources; nevertheless, the proposed IR model lacks of any weighting method, being a Boolean model. The documents are represented by the concepts associated to the instances in the document, but unlike the model of the present invention, the instances are not represented in the model. To rank the documents according to a user query, the distance function is extended among concepts to sets of concepts to define in this way a distance measure between documents. The distance between sets of concepts (documents) is defined as follows: given two documents or queries, its ranking distance is defined as the average minimum distance among all the pair wise combinations of concepts in the two sets. To achieve that the distance function on sets of concepts can verify the axioms for a metric, the distance function among sets of concepts is forced to be zero when the two input sets of concepts are equal. The last modification was defined to force the verification of the zero property axiom of a metric, but as result of it, the authors report undesired continuity problem near the zero distance value.

[0069] The IR model proposed in [Rada et al., 1989] is close to the IR model and method proposed in the present invention, which we call Intrinsic Ontological Spaces. Both models are the unique ontology-based IR models that use a metric space for the representation of the indexed information units. We can find some similarities and differences among both models in some aspects, such as: the use of ontology-based semantic distances, the representation of documents by sets (not vectors) of concepts, and the definition of a rank function between sets of concepts.

[0070] There are several similarities and differences between both models. First, both models represent documents by sets of concepts, although the Intrinsic Ontology Spaces also includes instances of concepts (individuals). Second, both models use a semantic distance defined on the ontology, but while Rada et al. use the shortest path length, the present invention uses a generalization of the Jiang-Conrath distance [Jiang & Conrath, 1997] designed to remove known drawbacks in the edge-counting family of semantic distances and the standard Jiang-Conrath distance. Third, Rada’s model uses the average distance among all cross-pairs of elements to define a metric among sets of concepts, whereas the present invention uses the standard Hausdorff distance as metric, with the advantage that the Hausdorff distance is better founded from a mathematical point of view [Henrikson, 1999]. Unlike the Rada’s distance among sets, the Hausdorff distance selects the maximum distance among all the point-set distance values. The Hausdorff distance is the induced metric on subsets of a metric space as result from the extension of the metric of the space to sets. The Hausdorff

distance is always continuous according to the topology induced by the metric of the space, removing the drawback related to the continuity around zero that Rada et al. report for their ranking function in [Rada et al., 1989].

[0071] VSM-based ontology-based IR models. The more recent family of ontology-based IR models start with the pioneering works in [Vallet et al., 2005] and [Fang et al., 2005]. Both works were independently published in very close dates, without any cross citation between them, or in others subsequent works as [Castells et al., 2007] and [Fernandez Sánchez, 2009]. We categorize these pioneering works, and all the subsequent works reviewed herein, in a subfamily of ontology-based IR models called vector-based conceptual models, or adapted VSM models, because all of them share a common approach based in the adaptation of a classical Vector Space Model (VSM) [Salton et al. 1975] to represent concepts, instead of keywords.

[0072] The IR model proposed in [Vallet et al., 2005] was continued in [Castells et al., 2007], being this research trend the core of the thesis of Miriam Fernández [Fernández Sánchez, 2009].

[0073] In [Vallet et al., 2005] and [Castells et al., 2007], the authors propose an ontology-based IR model based in the adaptation of a VSM model to represent concepts and individuals instead of meaningless terms. This model includes most part of the features exhibited by the models in the ontology-based IR family, and it could be considered as the canonical representative of this family.

[0074] The main idea in [Castells et al., 2007] is to substitute the keywords vocabulary of a classic keyword-based (KB) VSM, which defines the base vector set, for a vocabulary of concepts and instances within the base ontology of the KB, instead of a collection of meaningless terms. The documents are represented (indexed) by a vector of adapted TFIDF (Term Frequency—Inverse Document Frequency) weights, where each weight is defined according to the saliency of a concept, or instance of a concept, within a document, and its semantic discrimination capacity. Each document is represented by a set of concepts and concept instances, instead of keywords, in this way, the system index the documents using concepts and instances as base vectors of its VSM model. To index the documents, the system associates a set of semantic annotations for the found references to concepts in the knowledge base (KB), which define the collection of concepts instantiated within each document. The automatic semantic annotation is a very complex task which still is a very active research field in the information extraction (IE) community; for this reason, the automatic semantic annotation problem is out of the scope of the present invention, such as it is made in [Castells et al., 2007], and it is assumed that the IR model, method and system herein proposed need to be integrated with additional IE components for this task.

[0075] The operation of the IR model proposed in [Castells et al., 2007] is as follows. First, the system only accepts user queries in SPARQL format and it assumes that the documents have already been semantically annotated. Second, each document is represented by a set of semantic annotations within an ontology, which are defined by the references to concepts found in the documents. Third, the SPARQL query is used to interrogate the ontology and to retrieve all the documents with annotations derived from the concepts and instances included in the query. Fourth, all the documents retrieved are represented by vectors before to be ranked, while the base of the vector space is defined by all the con-

cepts and instances (individuals) included in the ontology, and an adaptation of TFIDF weighting scheme is used to convert the set of annotations of each document in a normalized vector expressed in the base of the concept-based vector space. Finally, the retrieved documents are ranked using the cosine function. The system is a direct and natural adaptation of the classic VSM model to manage concepts. The proposal in [Castells et al., 2007] agrees with other cited authors in that the proposed semantic IR model needs to be combined with standard keywords-based VSM models, due to the impossibility to have broad covering ontologies in a near future; for this reason, their system builds two independent VSM models (keywords and concepts) that are combined in the last retrieval stage.

[0076] The semantic retrieval capability of the Castells-Fernandez-Vallet model is derived from the semantic retrieval of annotated document in the ontology, which is able to retrieve documents with references to concepts not included in the query, starting from more abstract concepts provided in the query. This capability is the essential contribution derived from the use of ontologies in IR, as well as the main reason to its broad acceptance in all sort of semantic search applications.

[0077] The documents retrieved by the Castells-Fernandez-Vallet model are the documents annotated with entities found in the document collection retrieved by the SPARQL query [Castells et al., 2007], but the work does not clarify how it manages, if it does, the mentions to classes of concepts as mention to sets of subsumed concepts and instances. A mention to a class as set of elements means that the name of class is being used to retrieve all the information units that includes semantic annotation of concepts or instances subsumed by the name of class, in other words, the mention to a class is acting like a selection operator for the whole set of subsumed concepts.

[0078] The model in [Castells et al., 2007] was extended in [Fernandez Sánchez, 2009] and [Fernandez et al., 2011], broadening its application to a large scale and heterogeneous context as the web. Meanwhile, in [Bratsas et al., 2007], the authors introduce an application of the model in [Castells et al., 2007] to the problem of information retrieval in biomedicine, using a domain specific ontology and a fuzzy query expansion.

[0079] In [Fang et al., 2005], the authors propose an ontology-based IR model almost identical to the model in [Castells et al., 2007]. The model of Fang et al. has the same functional structure that the Castells-Fernández-Vallet model. The system admits queries defined by keywords or complex expressions which are transformed to queries in format OWL-DL. The OWL queries retrieve the related RDF triplets contained in the knowledge base (KB) with references to the concepts and instances included in the user query. From the concepts and instances in the RDF triplets, the system retrieves the associated documents, and lastly, the documents are ranked according to the user query using the cosine function. Such as in [Vallet et al., 2005], the model in [Fang et al., 2005] builds an adapted VSM representation through a TDIDF weighting scheme using the instances-document frequency matrix, but unlike [Vallet et al., 2005], the final weights include a saliency factor whose purpose is to take into account the semantic differences among concepts and instances.

[0080] The work of [Fang et al., 2005] can be considered as a first try to include a semantic distance measure in an ontology-based IR model, although it is a coarse approximation,

because the theory about ontology-based semantic distances offers a well-founded and precise solution to this problem. Precisely, the Intrinsic Ontological Spaces model builds on an extension of previous results on this theory, with the aim to provide a unified representation that integrates the intrinsic structures of the ontology in the model, providing many potential benefits to the users while it overcomes the common drawbacks of the family of ontology-based IR models.

[0081] In [Mustafa et al., 2008], the authors propose a semantic IR model based in the use of RDF triplets and a thematic similarity function. The thematic similarity function associates concepts according to its membership in a common semantic field or theme. The user queries are encoded as RDF triplets, which are expanded to include synonyms and other semantically related concepts. The query expansion with related concepts uses a neighbourhood notion based in a measure of semantic distance among concepts on the ontology. To establish the semantic similarity among the queries and the documents, the system uses the RDF triplets in the query and the RDF annotations associated to the documents. The documents with RDF triplets matching the terms in the expanded query are extracted from the collection, and are ranked according to their saliency. To select the documents that match the query terms, the authors use a set of semantic distance functions on the ontology to compute the closeness among the concepts in the query and the concepts annotated by the document, in other words, the retrieval of documents is driven by a ontology-based semantic distance function instead of a formal Boolean SPARQL query. To rank the retrieved documents, the documents are represented into a vector space of RDF concepts using a TFIDF weighting scheme; then, the documents are ranked using a combination of the cosine function and the same semantic distance function that was previously used. The semantic distance function used in the IR model is a novel edge-counting measure proposed in the same work, which includes an exponentially decreasing factor according to the depth of the nodes. The methodology can be summarized in four steps: (1) query expansion of the RDF triplets, (2) retrieval of related documents based in a novel edge-counting semantic distance, (3) mapping of the documents to a concept-based vector space using a TFIDF weighting, and (4) document ranking using the standard cosine function. The main drawback of the model of Mustafa et al. is that it retains the same geometric inconsistencies that previous ontology-based IR models, despite its smart integration of the semantic distances in the retrieval process. Although the model retrieves the documents using an ontology-based semantic distance, notion that we share as support in the model of the present invention, in [Mustafa et al., 2008] the documents are ranked in a concept-based vector space where the semantic metric is missing. A second drawback of the model is the use of an edge-counting distance, which have been refuted by the research community, such as it will be later discussed when explaining the ontology-based semantic distances. Today, the Jiang-Conrath distance [Jiang & Conrath, 1997], in combination with any IC-based intrinsic method to get the Information Content (IC) values, is one of the most broadly accepted semantic distances in the literature [Sánchez et al., 2012]. Lastly, another drawback of the IR model in [Mustafa et al., 2008] is that it does not consider instances of concepts, or named entities, in its representation, in contrast with the Intrinsic Ontological Spaces model proposed in the present invention.

[0082] In [Dragoni et al., 2010], the authors propose a concept-based vector space model which uses WordNet leaf concepts as base vectors for the representation of documents and queries. The proposed IR model is an adapted concept-based VSM model with an adapted TDIDF weighting method, and the standard cosine function as method for the ranking of saliency documents. The paper does not give details about the process to convert terms in WN concepts. Because the model does not include abstract concepts in its vocabulary, all the explicit references in the texts to abstract concepts not included in the vocabulary are discarded by the system. Also, the model does not include named entities recognizers (NER). Like the other concept-based adapted VSM models already described, the model of Dragoni et al. falls in the same modelling inconsistencies reported in the motivation section.

[0083] In [Egozi et al., 2011], the authors introduce a novel conceptual IR model based in the extension of a keywords-based VSM model with concepts defined in an ontological KB. Both, documents and queries are represented by a vector of weighted terms enriched with weighted concepts obtained through the use of an automatic annotation method, which extracts the underlying concepts within both text sources. The automatic annotation method used is called Explicit Semantic Analysis [Gabrilovich & Markovitch, 2006], and it is used to expand the standard terms-based VSM representation. The concepts used in the model are extracted from a hand-coded ontology. The authors use a feature selection method to choose the subset of concepts that best represents the corpus, and the selected concepts are used to expand the keywords-based VSM representation. The model proposed improves the results of previous methods when it is evaluated over some TREC corpus. By other hand, this model joins keywords and abstracts concepts in a same VSM model, falling into the cardinality mismatch problem reported above. The authors follow the idea mentioned in [Castells et al., 2007] about the use of the ontology-based models as a complement to standard keywords-based models. The Explicit Semantic Analysis (ESA) model does not use a formal ontology to describe the structure relations of the concepts, although it could be easily extended to do it, such as is made by the authors in their proposal.

[0084] Besides the common drawbacks of the family of ontology-based IR models previously described, the main drawback of the model in [Egozi et al., 2011] is that it only includes references to abstract concepts (classes), not to entities (instances). From an abstract point of view, the model of Egozi et al. uses the same strategies that the models in [Castells et al., 2007], [Fang et al., 2005] and [Mustafa et al., 2008]. These strategies can be summarized as follows: (1) use a concept-based representation for documents and queries, (2) the use of ontologies, and (3) indexing and retrieval of documents based in a concept-based adaptation of the VSM model.

[0085] Unlike the model in [Castells et al., 2007], which builds two independent vector representations (keywords-based and concepts-based) that are combined later in the retrieval stage, the model in [Egozi et al., 2011] mixes concepts and meaningless terms in the same VSM representation. Precisely, the core idea of the work in [Egozi et al., 2011] is to enrich the vocabulary based in keywords with concepts. The references to entities are captured by the meaningless keywords or terms, while the references to abstract concepts are captured through the ESA annotation method.

[0086] In [Cao & Ngo, 2012], the authors propose an extension of the keywords-based VSM model with ontological features associated to the named entities. The basic hypothesis is that the named entities are the more discriminative terms in most of the user queries; therefore, the enrichment of the VSM model with information not explicitly represented in the documents should lead to improvements in the precision and recall measures. The main idea is to merge in a same vector representation the TFIDF weights derived from independent vocabularies with features from different nature. The model uses a multivector representation for each document, where each document is defined by a vector of TFIDF weights defined on multiples vocabularies associated to the different types of features, such as: keywords, the alias, the associated class to the named entity, and entity identifiers among others. By last, the model uses a barycentric combination of the cosine function for each independent vector, such that the similarity between a document and a query is a weighted function of the individual similarities among pairs of independent feature vectors. The weight factors used to merge the independent similarity measures are left as free parameters to be tuned by each application. Again, this adapted VSM model falls in the same modelling inconsistencies already reported.

[0087] In [Machhour & Kassou, 2013], the authors introduce a method to integrate the use of ontologies in VSM-based systems for text categorization (TC) already existent. The core idea of the method is to map the original term-based vectors, whose coordinates represent meaningless terms, to concept-based vectors whose coordinates represent concepts within ontology. The authors evaluate the proposed model with the known RCV1 corpus [Lewis et al., 2004], reporting only small improvements in performance, which they attribute to the strong pre-processing of these systems (stemming without disambiguation). Despite these discouraging results, the work studies a practical open problem with a clear application in TC.

[0088] In patent document US 2008/0270384 A1 the authors disclose a system and method for intelligent ontology-based knowledge search engine, called IATOPIA KnowledgeSeeker, which introduces a concept-based clustering method and a semantic annotation method for Chinese web articles. It can be applied to search the web, as in the embodiment that they present for news articles, using ontologies to analyze the semantics of Chinese texts. The components (modules of the system and method) are described in a detailed way as follows: (1) the topic ontology to model the kind (topic/several topics) of the articles, (2) the article ontology to represent the semantic content of the articles, (3) the lexical ontology to “understand” the semantics of Chinese text in HowNet. The system indexes HTML web pages containing articles categorized in the topic ontology; then, the system extracts the semantic content in the articles using an automatic semantic annotation method based in the article and lexical ontologies. The system produces a set of RDF triples as semantic annotation of every indexed web page. RDF annotation enables semantic quering on the classes, attributes and properties defined or from imported ontologies. The news recommendation uses two approaches: a personalized content based recommendation that is based on user preferences. The article ontology allows the representation of the structure of the article (headline, abstract, body, etc.), as well as other metadata (author, date, organization, etc.). On other hand, the lexical ontology allows the annotation of the semantic content within the article through the identification

of the concepts associated to the Chinese words. The lexical ontology is based in a bilingual Chinese-English resource called HowNet. Every Topic class is represented by a vector of weighted-concepts, called “sememes”, whose weights are obtained from a corpus through a TFIDF weighting method. The proposed model uses a concept-based Vector Space Model (VSM), where the sememes are concept-based vectors with features derived from the lexical ontology. The indexing process is as follows: (1) one input article in HTML is semantically annotated by RDF triples using an automatic method and the article and lexical ontology, (2) the new indexed article is represented by a concept-based vector, whose features are ontology-based annotations, (3) the system computes the score similarity (cosine function) among the input article and the centroid vectors for each topic, and (4) the topic with the maximum score is selected to categorize the content of the article. The proposed method uses a concept-based VSM model to represent the indexed documents, and the cosine function as the score (ranking) function to carry-out the clustering of the input documents, thus, this method presents the same drawbacks and inconsistencies of the rest of the models in the adapted VSM models.

[0089] In patent document US 2014/0074826 A1 the authors disclose a novel vector ontology-based information retrieval (IR) system, which uses semantic annotations enriched with linguistic info, and a linear combination of multiple scores as document relevance function. The system parses any query into lexical elements defined by words or phrases; then, it computes automatically a semantic index composed by a set of ontology-based semantic annotations. Each semantic annotation is defined by a concept, or instance of a concept, and it is composed by the lexical token, concept, morphological invariant form (stem) of the token, and the part-of-speech. The semantic index of the query is used to retrieve and score the related indexed documents. The system ranks the retrieved documents using multiple scoring functions, while it uses the same language processing and semantic annotation for queries and the indexed documents. The system uses a query-matching language (IML) to analyze the user’s queries, and a rule engine converts the queries in a set of search criteria and into actions list. The response engine of the system takes the actions list and search criteria as input, and it defines a document-level relevance method based in a linear combination of four different scores as follows: (1) a TDFIF weighting term-based cosine score, (2) a concept-based vector cosine score, with a custom normalized weighting function, (3) a stem-based cosine score using the same concept-based weighting function, and (4) a link-based score method. In sum, the scoring system to define the document relevance with regard to the input query is using three different Vector Space Models (term-based, stem-based and concept-based) plus a link-based ranking function. This system exhibits the same drawbacks than the rest of concepts-based vector IR models already cited above.

[0090] Patent document US 2008/0288442 A1 discloses a rule-based method and its corresponding system to decide if a set of statements (RDF) according to a specified ontology has to be stored, indexed or not at all. Several indices can be produced using the set of rules, a set partitioned regarding two functionalities: one part to decide what kinds of indexes are needed (for the textual part of the RDF triplets) and a second one to process every new statement. The method can also be used to mark up the ontology with metadata containing information about what statements with textual data has to be send

to the storage, to some of the indexes or not. The main claim of this invention is that it helps to deal with the unorganized and unstructured web of data, by using RDF to represent this huge amount of information. The RDF statements can be stored very efficient and saving resources, but when the statements contain a textual part, the problem has to be addressed using indexes. That way the meaning of the query is also taking into account as the indexed terms are part of the RDF triplets and not isolate words. This method exhibits same drawbacks as the rest of approaches already cited in the family of VSM-based models.

[0091] Query-Expansion Ontology-Based IR Models.

[0092] The query-expansion ontology-based IR methods share some common features, such as:

[0093] (1) the semantic expansion of the queries to increase the retrieval capability of relevant document,

[0094] (2) the use of multiple-feature enriched semantic keys, which includes different types of semantic predicates, grammar role, and other linguistic information,

[0095] (3) a document retrieval based in two stages, a first selection of candidate documents, and a second stage of matching and ranking,

[0096] (4) the lack of a unified representation space to make the comparison and ranking of the indexed units,

[0097] (5) the lack of the use of ontology-based semantic distances as it is proposed in the present invention, and

[0098] (6) the use of ad-hoc scoring methods combining different semantic features.

[0099] In U.S. Pat. No. 6,675,159 B1 the authors disclose a concept-based indexing and search system for collections of documents, which have been semantically annotated with ontology-based predicates. The system extracts the concepts associated to any user query, and returns only those documents that match the concepts in the query. The documents and queries are represented by ontology-based predicates of different types, which encode “is-a” relations, verb arguments, or other semantic relations among individuals. The system uses an ontology-based parser to extract a collection of semantic predicates from the queries and documents, to be used as its representation. The retrieval of related documents is made in two stages: (1) candidate selection, and (2) comparison and ranking. The system holds the indexed corpus organized in domain-based clusters, and it uses a naive Bayes classifier to get the closest cluster for any input query. The filtered documents are semantically compared and ranked with regards to the query using a scoring method that combines weighted scores designed by type of semantic predicates. In spite of the fact that the documents are represented by a rich set of ontology-based predicates, the proposed method lacks from a unified representation space for the retrieval, comparison and ranking of documents. The lack of a unified representation space prevents the use of efficient search and ranking methods; moreover, the scoring function is not using a well-founded ontology-based semantic metric to compare concepts, such as the one that is proposed in this invention.

[0100] In U.S. Pat. No. 8,301,633 B2 the authors disclose a system and a method for the indexing and semantic search in a corpus of documents. The indexed documents are structured in passages, and the last ones are defined as the main indexed unit. The system builds an inverted index that relates every index term with all the passages where it appears, and every passage is related with every document. Every passage is represented by an inverted semantic key composed by differ-

ent fields associated to the index terms, called tokens, such as (a) the key term in a lexicon, (b) an ontology-based semantic annotation of the key term, (c) a semantic role labelling annotation, (d) the grammar role, (e) some linguistic annotations, and (f) some transformation rules. A key term is an index term defined in a key term lexicon, which represents an occurring word in an indexed document. Every index term can be the occurring word, or other related word, such as a synonymous, hypernym, or hyponym. The user queries are transformed to the semantic representation of the system, and the retrieval is carried-out in two stages, the search and the retrieval, as follows: (1) in the search stage, the system extracts a collection of candidate passages using the key term of the expanded query, and (2) in the retrieval stage, the system compute both, the semantic matching among the query and the passages, and the ranking using the full semantic representation for the query and the passages. The semantic matching of the queries and documents is a high computational task whether the full semantic representation is used, thus, the system uses a pre-selection stage based in a keyword-based query expansion method to retrieve all the candidate passages. The semantic representation of the input query is transformed into a collection of index terms using all sort of semantic relations according to the fields included in the semantic keys, by a query expansion method. The system retrieves all the passages matching any index term in the expanded query, then, all the retrieved documents are merged or discarded using a Boolean set of operations with the retrieved passages, according to the semantic representation of the query. Unlike the present invention, this system does not include any weighting method, because it makes an exhaustive Boolean keyword-based semantic annotation at the level of passage. The system does not use any sort of semantic space as mean for the indexing, retrieval and ranking of documents, but an exhaustive inverted index based in the semantic annotation of keywords defined in a lexicon. The final ranking score of the document combines multiple scores derived from the combination of the multiple fields in the semantic keys. One score is a semantic distance based in the use of the order relation in an ordered list of concepts, or keywords related to the index terms in the query. The passage-based indexing and the rich semantic and linguistic structures used as semantic keys for the passages are well-tuned in the context of passage retrieval for question-answering (Q&A) systems; however, the system presents some drawbacks in the context of a more general semantic search system for documents, or any sort of semantically annotated data. First, the high computational complexity derived from the indexing based in passages and the multiple-fields semantic keys. Second, the semantic score lacks of a well-founded ontology-based semantic metric as it is proposed in the present invention; thus, the intrinsic semantic distance among concepts is missing in the final ranking. Third, the lack of a unified and well-founded geometric representation space for the semantic keys prevents the system to be able to use efficient search algorithms for the comparison and ranking of documents.

2. Geometric Representation for Taxonomies

[0101] The present invention is related with one distance-preserving ontology embedding proposed by Clarke in [Clarke, 2007], whose main ideas has been also published in [Clarke, 2009] and [Clarke, 2012]. Following some geometric ideas introduced by Widdows in [Widdows, 2004], Clarke proposes a distance-preserving embedding method for the

concepts within a taxonomy, which is called vector lattice completion, whose main idea is to use the natural morphism between the taxonomies and the vector lattices.

[0102] Clarke's ideas are based in the very close relation between taxonomies and lattices, derived from the fact that many human-made taxonomies are join-semi lattices, although in the more general case, we could also find examples of taxonomies with multiple inheritance, where a pair of concepts do not have a supremum.

[0103] The vector completion builds an order preserving homomorphism which maps each concept to a linear subspace in the vector lattice, with the property that the Jiang-Conrath distance between concepts [Jiang & Conrath, 1997] is preserved as the Euclidean distance between vectors when the taxonomy is tree-like. The leaf concepts are mapped to base vectors of the space, while any non-leaf concept is mapped to the linear subspace spanned by its children concepts. The ontology embedding of Clarke is an implicit application of the theory of categories [Pierce, 1991], wherein his completion is a natural structure-preserving mapping among different, but intrinsically identical, algebraic structures.

[0104] Although the embedding proposed by Clarke represents a very important milestone in the search of a semantic distance-preserving representation for ontologies, and its application to the development of good ontology-based IR models, Clarke's work has two important drawbacks in the context of an ontology-based IR model that differentiate it with the model proposed here: (1) the lack of the integration of individuals (instances of concepts) in the model, and (2) the lack of the method to represent information units composed by a collection of concepts or references to them, such as documents. Unlike the model of the present invention, Clarke's embedding does not consider populated ontologies in his model, in other words, the vector lattice completion only works for concepts, not for individuals (instances). Moreover, the model of Clarke cannot be used to represent information units defined by a collection of concepts, or references to concepts (instances); in other words, we do not know how to use the vector lattice completion for representing and comparing documents. Precisely, Clarke surveys the compositionality vector-based representation problem in a recent work [Clarke, 2012].

3. Ontology-Based Semantic Distances

[0105] The necessity to compare semantic concepts has motivated the development of many semantic distances and similarity measures on ontologies. The distance and similarity functions are complementary functions with opposite meanings, in the sense that they produce antitone or opposite orderings, it means that for a greater similarity decreases the distance and vice versa. Any similarity function can be converted in a distance function, and vice versa; thus, we herein focus on the study of semantic distances on ontologies. For example, in the VSM model most of models employ the cosine function on the unit hypersphere (normalized vectors) which is exactly the opposite function of the geodesic distance among points on the feature space (unit n-sphere) of the model. The cosine function and the geodesic distance compute opposite orderings, but they produce exactly the same rankings of relevant documents for any input query.

[0106] An ontology-based semantic distance is a metric defined on the set of classes of any ontology, while an ontology-based semantic similarity is a similarity measure. We refer to both types of measures as ontology-based semantic

measures. The ontology-based semantic measures (distances and similarities) can be categorized in three broad classes:

[0107] (1) edge-counting based, such as the measures proposed in [Rada et al., 1989], [Lee et al., 1993], [Wu & Palmer, 1994] and [Hirst & St-Onge, 1998];

[0108] (2) vector-based, such as the measures proposed in [Frakes & Baeza-Yates, 1992]; and

[0109] (3), IC-based (IC stands for “Information Content”), whose main references are the proposals in [Resnik, 1995], [Jiang & Conrath, 1997] and [Lin, 1998].

[0110] The most broadly accepted family of measures are based in the information content (IC) of the concepts within a taxonomy. The IC-based family is subdivided in two sub-groups: (a) corpus-based methods, which use corpus-based statistics to compute the occurrence probabilities and the IC values for each concept, and (b) the intrinsic methods which only use the information encoded in the structure of the ontology, in whose family we can cite the pioneering works of [Seco et al., 2004] and [Zhou et al., 2008].

[0111] Any IC-based semantic measure is the combination of two complementary methods: (1) the measure function between concepts, properly called as IC-based measure, and (2) the method used to compute the IC values for the taxonomy’s nodes, called as IC-computation method. Thus, any IC-based semantic measure can be combined with any independent IC-computation method. By example, the Jiang-Conrath distance can be combined with any intrinsic IC-computation as the described ones in [Seco et al., 2004] and [Zhou et al., 2008].

[0112] The state of the art in semantic distances is defined by the IC-based measures disclosed in [Sánchez et al., 2012] and [Meng et al., 2012]. The main research trend in this area is the development of intrinsic IC-computation methods which use the intrinsic knowledge encoded in the ontology as means to remove the necessity to compute corpus-based statistics, as well as novel IC-based measures. The research activity in intrinsic IC-based methods has increased very recently.

[0113] According to some relevant benchmarks driven in the literature, we can conclude that the Jiang-Conrath distance offers some of the best results for most of the applications, in special, whether its IC values are estimated by any intrinsic method. In [Budanitsky & Hirst, 2001], the authors carry-out some benchmarks to compare the IC-based measures of Resnik, Jiang-Conrath, Leacock-Chodorow, Lin and Hirst-St-Onge, concluding that the Jiang-Conrath (JC) distance offers the best results. In a later work [Budanitsky & Hirst, 2006], the same authors arrive to the same conclusion, and the work includes cites to other reports with similar conclusions about the JC distance. In [Sánchez et al., 2011] the authors carry-out a benchmark among IC-based measures comparing corpus-based methods versus methods based on the computation of the IC values through intrinsic method. This last report concludes that all the measures work better using intrinsic IC computation, while the intrinsic Jiang-Conrath distance gets the second best global results for their tests, with a tiny difference to the first one. Most of the main benchmarks for ontology-based semantic measures consist in the evaluation of the semantic similarity between word pairs within the Wordnet [Miller, 1995] taxonomy.

[0114] We herein only survey the most representative measures in the cited categories. For a broader revision of the literature, we refer to some recent surveys, some of them are focused in biomedicine, such as [Lord et al., 2003], [Lee et al.,

2008], [Pesquita et al., 2009], [Hsieh et al., 2013i], [Cross et al., 2013], and [Harispe et al., 2014], while others do not assume any specific domain, such is the case in [Saruladha et al., 2010], [Sánchez et al., 2012], [Xu & Shi, 2012], and [Gan et al., 2013]. The book by Deza and Deza also includes a short, but very useful section about network-based semantic distances on ontologies as the Wordnet [Deza & Deza, 2009, sec. 22.2].

[0115] The first ontology-based semantic distances to appear were the edge-counting based measures, whose main representative is the Rada’s measure [Rada et al., 1989]. All these measures are characterized by the use of the shortest path length among concepts measured on the ontology graph. The key idea behind these methods is that the higher up you need to climb to find a common ancestor to both concepts, the greater should be the distance between concepts, and vice versa.

[0116] In [Rada et al., 1989], the authors propose to use the shortest path length among concepts of an ontology as distance measurement among them, measure that they call “distance”. Their work sets the first known ontology-based semantic distance, and it also introduces the main hypothesis underlying all the subsequent ontology-based semantic distances: the conceptual distance as metrics hypothesis. This hypothesis states, following previous psychological studies, that the conceptual distance, or similarity, among concepts in a semantic network, is proportional to the path length that joins them. The shortest path length, also called geodesic distance, is a metric in the formal sense; for this reason the authors in [Rada et al., 1989] prove that these measures are metrics on ontologies.

[0117] Other measures in the edge-counting family, such as the works in [Lee et al., 1993], [Wu & Palmer, 1994], [Leacock & Chodorow, 1998] and [Hirst & St-Onge, 1998], are also based in some combination of the shortest path values, as it can be appreciated in FIG. 3, and all of them share the same drawbacks.

[0118] FIG. 3 shows a summary of the formulas used by some known measures to compute the semantic similarity or distance between a pair of concepts within an ontology, as well as the novel distance disclosed in the present invention. The similarities appears as $\text{sim}(c_1, c_2)$ and distance functions as $d(c_1, c_2)$. The function $\text{de}(c_i)$ returns the depth of any concept in the direct acyclic graph (DAG) defined by the ontology, it means the length from the concept to the root node. By other hand, function $L(c_1, c_2)$ denotes the shortest path length among two concepts.

[0119] The main drawback of the measures based in edge-counting is that they implicitly assume that every edge has the same relevance in the computation of the global path length, without to take into account its depth level or occurrence probability. This drawback can be called the uniform weighting premise. In [Resnik, 1995], the authors propose a new semantic distance based in an Information Content (IC) measure whose main motivation is to remove the uniform weighting premise of the edge-counting measures. The IC measure for every concept is only the negative logarithm of the occurrence probability of the concept, such as is shown in equation (1), information content for every node within the taxonomy, defining a probability space, whose integral value on the ontology is 1. Resnik et al. define a similarity measure shown in FIG. 3, which is equivalent to assign a weight with the value of the probability difference between the adjacent concepts of each edge.

$$\begin{cases} p: C \rightarrow [0, 1] \subset \mathbb{R} \\ IC(c_i) = -\log_2(p(c_i)) \end{cases} \quad (1)$$

[0120] The key idea behind the IC-based distances is as follows. The probability function p in equation (1) is growing monotone while the ontology is bottom-up; thus, while we climb on the ontology, the observation probability of any abstract concept increases. As higher is the occurrence probability of one concept, lower is its information content and vice versa.

[0121] In [Jiang & Conrath, 1997], the authors propose a set of IC-based semantic distances encoding a set of semantics notions that fill some gaps in [Resnik, 1995]. Jiang and Conrath follow the IC approach of Resnik, but they note that previous measures not consider some important semantic notions encoded by an ontology, which affects the semantic similarity appreciated by the human beings. They consider the following issues: the number of descendants, the global depth of the concepts, the type of semantic relation (hyper/hypo/meronym), and the strength degree of a link between a parent concept and its children concepts. From the different measures proposed in [Jiang & Conrath, 1997], the more broadly adopted, also known as the JC measure, is the distance shown in FIG. 3.

[0122] In [Lin, 1998], the author refutes the vector-based distances, such as the proposed in [Frakes & Baeza-Yates, 1992], by the necessity to use vectors. Moreover, Lin also notes that the edge counting methods only works on taxonomies, not admitting other ways of knowledge representation, such as first order logic. Lin proposes a novel definition of semantic similarity based on a probabilistic model and the IC value.

[0123] The semantic distance proposed in [Jiang & Conrath, 1997] has three drawbacks that are solved by the novel ontology-based semantic distance proposed by the present invention, which is called weighted Jiang-Conrath. These drawbacks are the following:

[0124] (1) the Jiang-Conrath distance is only a metric in a strict sense when the ontology is tree-like, therefore the Jiang-Conrath does not satisfy the metric axioms on ontologies with lattice or general poset structure [Orum & Joslyn, 2009];

[0125] (2) the Jiang-Conrath distance is only uniquely defined for upper semi-lattice ontologies, not for ontologies with lattice or general poset structure; and

[0126] (3) it is only defined on taxonomies of concepts, not weighted concepts (classes) or instances of concepts (individuals).

[0127] The standard formula of the Jiang-Conrath distance on taxonomies is given by equation (2), where the term LCA (c_1, c_2) means the lowest common ancestor node between the concepts c_1, c_2 , and it could be written as $c_1 \vee c_2$ when the taxonomy is a join semi-lattice, because in this case every pair of concepts holds a supremum element.

$$d(c_1, c_2) = IC(c_1) + IC(c_2) - 2IC(LCA(c_1, c_2)) \quad (2)$$

[0128] Equation (2) is uniquely defined for lattices, being the upper semi-lattice taxonomies a special case. Any lattice is by definition a partially ordered set (poset) where each pair of elements shares a unique lowest common ancestor, called supremum. Therefore if the ontology is an upper semi-lattice, we find that any pair of concepts shares a unique common ancestor, and the third term in the equation (2) is well defined.

By contrast, for general taxonomies that not fulfil the lattice axioms, we find pairs of concepts with more than one lowest common ancestor.

[0129] Taxonomies can be classified in three classes according to its structure, such as it is shown below. These three classes of taxonomies define a hierarchy of sets, in the sense that the set of general taxonomies subsumes the set of lattice taxonomies, and the last one subsumes the set of the tree-like taxonomies.

[0130] (1) Tree-like taxonomies (see FIGS. 4 and 6). FIG. 4 represents an example of a tree-like ontology, which is a partial sub-graph of WordNet around the “armchair” concept. FIG. 6 represents a tree-like ontology with the edge weights defined as the difference of Information Content (IC) values between its extreme nodes. These edge weights match the implicit edge weights defined by the Jiang-Conrath distance.

[0131] (2) Upper semi lattices (see FIG. 7). In FIG. 7 every pair of concepts has a unique lowest common ancestor, called supremum in this case. The taxonomy exhibits a structured type of multiple inheritance that verifies the semi-lattice axioms.

[0132] (3) General partially ordered sets (as the ones shown in FIGS. 5 and 8). In particular, FIG. 5 represents the complete lattice associated to the power set (i.e. set of all subsets) for the set $\{1,2,3\}$. FIG. 8 represents a taxonomy with the structure of a general partially ordered set (poset structure). In this case, the taxonomy exhibits an unstructured multiple inheritance. The concepts “m” and “p” do not have a supremum, because they share two lowest common ancestors, the concepts “f” and “g”.

[0133] Starting from the observations above, we now summarize some of the main proven facts about the Jiang-Conrath distance. First, up to date, the Jiang-Conrath distance has proved to offer likely the best results for a semantic similarity/distance measure. This conclusion rises from many benchmarks carried-out in the literature, among we can cite the works in [Budanitsky & Hirst, 2006] and [Sánchez et al., 2012]. Today, the state of the art is based in intrinsic IC-based measures, in special, some intrinsic variants of the JC measure, such as is reported in [Sánchez et al., 2012].

[0134] Second, the Jiang-Conrath distance is uniquely defined only for taxonomies that verify the upper semi-lattice structure, such as the trees. Such as we explained above, this property is a consequence of the definition of the term $IC(LCA(c_1, c_2))$ as a function of the lowest common ancestor.

[0135] Third, the Jiang-Conrath distance is strictly a metric only on trees, not on semi lattices or general posets. The Jiang-Conrath distance is only uniquely defined on semi lattices where every pair of nodes has a unique supremum, or unique Lowest Common Ancestor (LCA); however, in [Orum & Joslyn, 2009], the authors prove that this condition is not enough to verify the axioms for a metric, because for some general rooted-posets (taxonomies) can happen that the triangle inequality was not satisfied. This theoretical result contradicts the claim made by Jiang and Conrath in their original paper [Jiang & Conrath, 1997], where they claim that their distance is a metric on any sort of taxonomy, without to include any exhaustive formal proof with regard it.

[0136] Fourth, the Jiang-Conrath distance is not uniquely defined on general taxonomies. For the case of general posets, the JC distance not only is not a metric, not even is well defined. The reason is that in this general case the existence of pairs of concepts with more than one LCA concept is possible. In a practical application, we can always select the first

LCA concept found in a LCA search, but we conjecture that it can produce discontinuities of distance function near of these elements, such as the discontinuity problems reported in [Rada et al., 1989] as consequence of the constraint imposed in their distance function among sets of concepts.

[0137] Fifth, the theoretical limitations of the Jiang-Conrath distance prevent to get a well-founded metric space on general taxonomies. One possible solution for the non-uniqueness condition would be to compute all the LCA values for each pair of concepts [Baumgart et al., 2006]; then, we could select the ancestral path with the minimum distance value as the Jiang-Conrath distance. This idea allows defining uniquely Jiang-Conrath distance on any taxonomy; nevertheless, it is not enough to verify the metric axioms, because, such as is proven in [Orum & Joslyn, 2009], it is not even possible in the simpler case of semi lattices, where the uniqueness condition is already guaranteed.

[0138] Sixth, the Jiang-Conrath distance between one concept and its parent is equal to the difference of their information content values. It means that any taxonomy endowed with the JC distance can be interpreted as a weighted-graph where each edge is weighted by the IC difference between its adjacent concepts.

[0139] Seventh, the JC distance between one concept and its parent is proportional to their joint probability. This fact is proven in [Jiang & Conrath, 1997], and it can be easily deduced.

[0140] The drawbacks of the Jiang-Conrath reported above motivate the development of the novel weighted Jiang-Conrath distance introduced in the present invention.

[0141] Intrinsic IC-based semantic distances. Today, it is broadly accepted by the research community that the IC-based semantic distance and similarities offer the best expected results in most of semantic evaluation tasks; however, the traditional IC-based family of methods has an important drawback from a practical point of view. The standard IC-based measures need to compute corpus-based statistics to evaluate the IC values for every concept within the ontology. The common method is to count every reference to a child concept as a reference to all its ancestors, and then using this frequency information to compute the occurrence probability for each concept on the ontology. The main problem with these corpus-based statistics is the difficulty to get well balanced corpus covering every concept in the ontology.

[0142] Motivated by the previous limitation, many authors have proposed novel methods, called intrinsic IC-based measures, whose main idea is to compute the IC values using only the information encoded in the same ontology, such as the density of the descendant nodes or its depth level respect to the root node.

[0143] The intrinsic methods are called IC-computation methods because they focus in the computation of the IC values used in combination with any IC-based semantic measure. It means that the intrinsic IC-computation is a complementary research problem associated to the development of ontology-based IC semantic measures. As pioneering works of this family, we can cite the works in [Seco et al., 2004] and [Zhou et al., 2008] and [Pirró, 2009].

[0144] The number of intrinsic IC-computation methods and intrinsic IC-measures proposed has grown rapidly during the last five years, converting the area in the main research trend in semantic distance and similarity measures. Among the collection of novel proposals, we can cite the works in [Pirró & Euzenat, 2010], [KhounSiavash & Baraani-

Dastjerdi, 2010], [Saruladha et al., 2011], [Sánchez et al., 2011], [Sánchez & Batet, 2012], [Taieb et al., 2012], [Lingling & Junzhong, 2012], [Cross et al., 2013], [Harispe et al., 2013] and [Gupta & Gautam, 2014]. In spite of this huge research activity, the only available survey on the cited topic is the work in [Meng et al., 2012], although it is already out of date.

[0145] The current IC-computation methods in the literature do not fulfil the structural constraints already described above as motivation for the development of a new set of edge-based IC-computation methods disclosed in this invention.

[0146] For the ontology-based IR model proposed in the present invention, any intrinsic node-based IC-computation method may be used to compute the IC-values for each node, such as the methods proposed in [Seco et al., 2004] and [Zhou et al., 2008], or any intrinsic edge-based IC-computation method, like the ones disclosed in the present invention, which will be later discussed, which allow the direct computation of the of edge weights. The preferred method for the edge-based IC-weights in the present invention is defined by the equation (8) below, and the edge-based IC value given in the equation (3), which is simply the negative binary logarithm of the joint probability $P(\text{child}|\text{parent})$.

[0147] To summarize the state of the art, all the ontology-based IR models revised fall in the category of concept-based adapted VSM models, with the exception of the model proposed in [Rada et al., 1989], which is based in the use of semantic metric spaces defined by one ontology-based semantic distance. The model of Rada et al. is close to the proposed Intrinsic Ontological Spaces model. Despite the great advances and results obtained by the family of ontology-based adapted VSM models, whose main representatives are the models of [Fang et al., 2005], [Castells et al., 2007] and [Mustafa et al., 2008], the ontology-based IR models can be improved if the modelling inconsistencies shared by these models are solved, as it is made by the present invention.

[0148] As previously discussed, despite the many semantic measures in the literature, it is broadly accepted that the Jiang-Conrath semantic distance offers some of the best results for most of the evaluated applications. The state of the art considers the use of the Jiang-Conrath distance measurement with some type of intrinsic IC estimation, such as the methods proposed in [Seco et al., 2004] and [Zhou et al., 2008]. The current research trend in semantic distance measurement is to develop novel intrinsic IC-based estimation methods and measurements. Moreover, the Jiang-Conrath is very well founded due to its connection to the lattice theory, and it defines a metric when the underlying ontology/taxonomy is tree-like, fact proven in [Orum & Joslyn, 2009].

[0149] [Clarke, 2007] proposes a distance-preserving embedding method for the concepts within a taxonomy, which is called vector lattice completion, whose main idea is to use the natural morphism between the taxonomies and the vector lattices. Because most of taxonomies fulfil the join semi lattice axioms, the ideal completion builds an order preserving homomorphism which maps each concept to a linear subspace in the vector lattice, with the property that the Jiang-Conrath distance among concepts is preserved as the Euclidean distance between vectors when the taxonomy is a tree. The leaf concepts are mapped to base vectors of the space, while any non-leaf concept is mapped to the linear subspace spanned by its children concepts. The ontology embedding of Clarke is an implicit application of the theory

of categories [Pierce, 1991], where his ideal completion is a natural structure-preserving mapping among different, but intrinsically identical, algebraic structures. Despite the fact that Clarke's model is not defined for individuals, it establishes a very important theoretical result: it is proven that any taxonomy can be embedded in a vector lattice, in such way that its topological structure (order) and metric structure (semantic distance) is preserved.

[0150] Next, a summary of the differences between the proposed method for the definition of an ontology-based IR model and the methods reported in the literature is provided.

[0151] First, unlike most of the previous methods, the present method represents the information units by sets of weighted-mentions to concepts (classes) or instances of concepts (individuals), instead of vectors whose coordinates represent weighted mentions on a set of mutually orthogonal vectors defined by the a set of concepts (classes) and/or instances of concepts (individuals).

[0152] Second, in the present invention the mentions to concepts (ontological classes) are represented by sets with the following structure. Every set of elements in the representation space, associated to the embedding of any class in the ontology, verifies the next property: the set subsumes all the subsets associated to the descendant classes (concept) and individuals (instance of concept) within the populated ontology, according to the metric space. It is the first time that a concept in the query is equivalent to the selection of a geometric subset of the representation space, that is, any logic query is converted in the selection of the geometric region containing all the concepts (classes) and instances (individuals) subsumed by the concept cited in the query.

[0153] Third, unlike other known methods, the present method integrates in the same semantic representation space the mentions to concepts (classes) and instances of concepts (individuals) in a consistent way, through the preservation of the structures defined by the intrinsic geometry of the base ontology.

[0154] Fourth, the present method explicitly integrates and preserves the intrinsic geometry of any base ontology in the representation space, given by the next structure relations: (1) the order relation of the taxonomy, (2) its intrinsic semantic distance, and (3) the set inclusion for the individuals and subsumed concepts of the ontology.

[0155] Fifth, the weighted-mentions to concepts or instances of concepts are represented in a metric space based in a novel ontology-based semantic distance, in contrast with most of methods that uses a vector space model (VSM) and the cosine function as similarity measure. The approach of the present invention removes the implicit orthogonality condition derived from the use of the cosine function as ranking method in every VSM-based ontology-based IR model in the literature, which is a source of semantic inconsistency in the previous representations.

[0156] Sixth, unlike other previous methods, the present method uses the Hausdorff distance as a metric on subsets of a metric space to compare and to rank information units (documents), instead of the cosine score. This feature also contributes to remove the implicit orthogonality condition of the VSM models. By other hand, the Hausdorff distance is well defined metric on subsets of a metric space, which allows to remove the continuity problems reported in [Rada et al., 1989], and to build a semantic ranking function supported by a meaningful ontology-based distance, such as the novel distance introduced in the present method.

[0157] Seventh, the proposed novel weighting method is defined as a statistical fingerprint, but it has a semantic meaning. The weight factor is a statistical and static weight derived from the frequency of every mention to a concept or instance within an information unit, equivalent to the standard TF (Term Frequency) weights used in all known IR models. However, the weight defines the ontology-based edge weight for each weighted-mention in the model, and it is a semantic weight defined by the IC-value of the mentioned ontological object. The novel weighting method proposed combines, for the first time, a statistical and static weight with an ontology-based semantic distance.

[0158] Eighth, the only known method that also uses a metric space for the representation of the information units is the model introduced in [Rada et al., 1989], but it presents some important differences respect to the present method. Firstly, the model of Rada et al. represents every document as a set of Boolean mentions to concepts, while our method includes a weighting method to represent the information units (documents) as a set of weighted-mentions to concepts and instances of concepts. Secondly, the model of Rada et al. uses the average ontology-based distance among concepts as a distance function among sets, while the present invention uses the Hausdorff distance, which is a strict metric among subsets and removes some continuity problems reported by the authors in [Rada et al., 1989]. Thirdly, the ontology-based distance of Rada et al. does not include the distance between instances of concepts in its model, and it is based in the shortest path distance between concepts, while the present method uses the shortest weighted path distance among concepts with weights defined by the novel weighted Jiang-Conrath distance.

[0159] Ninth, the present method proposes a novel ontology-based semantic distance based in the shortest weighted path on the populated ontology, where the weights are the negative logarithm of the joint probability between a child concept and its parent concept. The novel semantic distance is a generalization of the Jiang-Conrath distance, whose purpose is to remove the drawbacks described above. Unlike the standard Jiang-Conrath distance, the present method is a well-defined metric on any sort of ontology, while the first one is only a well-defined metric on tree-like ontologies.

[0160] Tenth, unlike the previous intrinsic IC-computations methods reported in the literature, our novel family of intrinsic IC-computation methods (IC-JointProbUniform, IC-JointProbHypo and IC-JointProbLeaves) fulfil the structural constraints relating the Information Content values, the joint probabilities and the underlying base taxonomy.

[0161] Eleventh, unlike the previous methods, the present method defines a novel IR model where each one of its components is ontology-based, avoiding the loss of any semantic information derived from the base ontology of the indexing model. First, the representation space is defined by a metric space of weighted-mentions to concepts and instances, whose metric is ontology-based. Second, the weighting method, in spite to be a classical TF scheme, has a semantic contribution to the distance among items in the populated ontology, because the weights define the joint probability for the weighted elements, whose IC-value is the length of edge joining any weighted-item to its parent concept/individual; thus, the weighting method is also ontology-based. Third, the ranking method is also ontology-based because it is based in the Hausdorff metric on subsets of the representing space, which derives directly from the ontology-based metric of the

space. Fourth, the retrieval method is driven by the ranking method, thus, the retrieval operation is also ontology-based. Fifth, the information units are represented by a set of weighted-mentions to individuals and classes within ontology; therefore, the representation is directly defined on the underlying populated ontology space plus a metric derived from its structure. Sixth, the retrieval and ranking process is directly carried-out using the representation of information units, which avoids the necessity to interrogate the populated ontology through any formal query in SPARQL, or other equivalent language.

BIBLIOGRAPHIC REFERENCES

- [0162] [Ahuja et al., 1990] Ahuja, R. K., Mehlhorn, K., Orlin, J., & Tarjan, R. E. (1990). Faster Algorithms for the Shortest Path Problem. *Journal of the ACM*, 37(2), 213-223.
- [0163] [Basin & Pennacchiotti, 2010] Basili, R. & Pennacchiotti, M. (2010). Distributional lexical semantics: Toward uniform representation paradigms for advanced acquisition and processing tasks. *Natural Language Engineering*, 16, pp. 347-358.
- [0164] [Baumgart et al., 2006] Baumgart, M., Eckhardt, S., Griebisch, J., Kosub, S., & Nowak, J. (2006). All-Pairs Common-Ancestor Problems in Weighted Dags. Technical Report TUM-I0606, Institut für Informatik, Technische Universität München.
- [0165] [Bratsas et al., 2007] Bratsas, C., Koutkias, V., Kaimakamis, E., Bamidis, P., & Maglaveras, N. (2007). Ontology-based vector space model and fuzzy query expansion to retrieve knowledge on medical computational problem solutions. In *Engineering in Medicine and Biology Society, 2007. EMBS 2007. 29th Annual International Conference of the IEEE* (pp. 3794-3797). IEEE.
- [0166] [Brin et al., 1995] Brin, S. (1995). Near neighbor search in large metric spaces. In *Proceedings of the 21st Conference on Very Large Databases (VLDB.95)* (pp. 574-584).
- [0167] [Budanitsky & Hirst, 2001] Budanitsky, A. & Hirst, G. (2001). Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. In *Workshop on WordNet and Other Lexical Resources*, volume 2.
- [0168] [Budanitsky & Hirst, 2006] Budanitsky, A. & Hirst, G. (2006). Evaluating wordnet-based measures of lexical semantic relatedness. *Computational Linguistics*, 32, 13-47.
- [0169] [Cao & Ngo, 2012] Cao, T. H. & Ngo, V. M. (2012). Semantic search by latent ontological features. *New Generation Computing*, 30, 53-71.
- [0170] [Castells, 2008] Castells, P. (2008). Búsqueda semántica basada en el conocimiento del dominio. In F. Verdejo & A. Garcia-Serrano (Eds.), *Acceso y visibilidad de la información en la red: el rol de la semántica* (pp. 111-138). España: Universidad Nacional de Educación a Distancia (UNED).
- [0171] [Castells et al., 2007] Castells, P., Fernández, M., & Vallet, D. (2007). An adaptation of the vector-space model for ontology-based information retrieval. *IEEE Trans. Knowl. Data Eng.*, 19(2), 261-272.
- [0172] [Chu-Carroll et al., 2012] Chu-Carroll, J., Fan, J., Boguraev, B., Cannel, D., Sheinwald, D., & Welty, C. (2012). Finding needles in the haystack: Search and candidate generation. *IBM J. Res. Dev.*, 56, 6: 1-6: 12.
- [0173] [Clark, 2012] Clark, S. (2012). Vector space models of lexical meaning. In S. Lappin & C. Fox (Eds.), *Handbook of Contemporary Semantics*. Malden, Mass.: Blackwell, second edition.
- [0174] [Clarke, 2007] Clarke, D. (2007). Context-theoretic Semantics for Natural Language: an Algebraic Framework. PhD thesis, University of Sussex.
- [0175] [Clarke, 2009] Clarke, D. (2009). Context-theoretic semantics for natural language: an overview. In R. Basili & M. Pennacchiotti (Eds.), *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics* (pp. 112-119). Athens, Greece: Association for Computational Linguistics.
- [0176] [Clarke, 2012] Clarke, D. (2012). A context-theoretic framework for compositionality in distributional semantics. *Comput. Linguist.*, 38, 41-71.
- [0177] [Cross et al., 2013] Cross, V., Yu, X., & Hu, X. (2013). Unifying ontological similarity measures: A theoretical and empirical investigation. *International Journal of Approximate Reasoning*, 54(7), 861-875.
- [0178] [Deza & Deza, 2009] Deza, M. & Deza, E. (2009). *Encyclopedia of distances*. Springer.
- [0179] [Ding et al., 2007] Ding, L., Kolari, P., Ding, Z., & Avancha, S. (2007). Using ontologies in the semantic web: A survey. In R. Sharman, R. Kishore, & R. Ramesh (Eds.), *Ontologies*, volume 14 of *Integrated Series in Information Systems* (pp. 5 79.113): Springer US.
- [0180] [Dragoni et al., 2010] Dragoni, M., Pereira, C. D. C., & Tettamanzi, A. G. (2010). An ontological representation of documents and queries for information retrieval systems. In *Trends in Applied Intelligent Systems* (pp. 555-564): Springer.
- [0181] [Egozi et al., 2011] Egozi, O., Markovitch, S., & Gabrilovich, E. (2011). Concept-based information retrieval using explicit semantic analysis. *ACM Transactions on Information Systems (TOIS)*, 29, 8.
- [0182] [Eilenberg & MacLane, 1945] Eilenberg, S. & MacLane, S. (1945). General theory of natural equivalences. *Trans. Amer. Math. Soc.*, 58, 231-294.
- [0183] [Erk, 2012] Erk, K. (2012). Vector space models of word meaning and phrase meaning: A survey. *Lang. Linguist. Compass*, 6, 635-653.
- [0184] [Erkan & Radev, 2004] Erkan, G. & Radev, D. R. (2004). LexRank: Graph-based lexical centrality as salience in text summarization. *J. Artif. Intell. Res.*, 22, 457-479.
- [0185] [Fang et al., 2005] Fang, W.-D., Zhang, L., Wang, Y.-X., & Dong, S.-B. (2005). Toward a semantic search engine based on ontologies. In *Proceedings of International Conference on Machine Learning and Cybernetics*, volume 3 (pp. 1913-1918). IEEE.
- [0186] [Fernández et al., 2011] Fernández, M., Cantador, I., López, V., Vallet, D., Castells, P., & Motta, E. (2011). Semantically enhanced information retrieval: An ontology-based approach. *JWS special issue on semantic search. Web Semantics: Science, Services and Agents on the World Wide Web*, 9(4), 434-452.
- [0187] [Fernández Sánchez, 2009] Fernández Sánchez, M. (2009). Semantically enhanced Information Retrieval: an ontology-based approach. PhD thesis, Universidad Autónoma de Madrid, Departamento de Ingeniería Informática.

- [0188] [Frakes & Baeza-Yates, 1992] Frakes, W. & Baeza-Yates, R. (1992). *Information Retrieval: Data Structures and Algorithms*. Prentice Hall PTR.
- [0189] [Gabrilovich & Markovitch, 2006] Gabrilovich, E. & Markovitch, S. (2006). Overcoming the brittleness bottleneck using wikipedia: Enhancing text categorization with encyclopaedic knowledge. In *Proceedings of the Twenty-First National Conference on Artificial Intelligence (AAAI 2006)*, volume 6 (pp. 1301-1306). Boston, USA: AAAI Press.
- [0190] [Gan et al., 2013] Gan, M., Dou, X., & Jiang, R. (2013). From ontology to semantic similarity: calculation of ontology-based semantic similarity. *Scientific World Journal*, 2013, 11.
- [0191] [García-Hernández & Ledeneva, 2009] García-Hernández, R. A. & Ledeneva, Y. (2009). Word sequence models for single text summarization. In *Second International Conference on Advances in Computer-Human Interactions (ACHI.09)* (pp. 44-48). IEEE.
- [0192] [Gold & Angel, 2006] Gold, C. & Angel, P. (2006). Voronoi hierarchies. In M. Raubal, H. Miller, A. Frank, & M. Goodchild (Eds.), *Geographic Information Science*, volume 4197 of *Lecture Notes in Computer Science* (pp. 99-111). Springer Berlin Heidelberg.
- [0193] [Gupta & Gautam, 2014] Gupta, A. & Gautam, K. (2014). Semantic similarity 5 measure using information content approach with depth for similarity calculation. *International Journal of Scientific & Technology Research*, 3(2), 165-169.
- [0194] [Harispe et al., 2014] Harispe, S., Sánchez, D., Ranwez, S., Janaqi, S., & Montmain, J. (2014). A framework for unifying ontology-based semantic similarity measures: A study in the biomedical domain. *Journal of Biomedical Informatics*, 48, 38-53.
- [0195] [Hatzivassiloglou et al., 2001] Hatzivassiloglou, V., Klavans, J. L., Holcombe, M. L., Barzilay, R., Kan, M.-Y., & McKeown, K. R. (2001). SimFinder: A flexible clustering tool for summarization. In *Proceedings of the NAACL workshop on automatic summarization* (pp. 41-49).
- [0196] [Henrikson, 1999] Henrikson, J. (1999). Completeness and total boundedness of the Hausdorff metric. *MIT Undergraduate Journal of Mathematics*.
- [0197] [Hirst & St-Onge, 1998] Hirst, G. & St-Onge, D. (1998). Lexical chains as representations of context for the detection and correction of malapropisms. In C. Fellbaum (Ed.), *WordNet: An electronic lexical database* (pp. 305-332).: Massachusetts Institute of Technology.
- [0198] [Hsieh et al., 2013] Hsieh, S.-L., Chang, W.-Y., Chen, C.-H., & Weng, Y. C. (2013). Semantic 20 similarity measures in the biomedical domain by leveraging a web search engine. *Biomedical and Health Informatics, IEEE Journal of*, 17(4), 853-861.
- [0199] [Jiang & Conrath, 1997] Jiang, J. J. & Conrath, D. W. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of International Conference Research on Computational Linguistics (ROCLING X)*.
- [0200] [Kannan et al., 2012] Kannan, P., Bala, P. S., & Aghila, G. (2012). A comparative study of multimedia retrieval using ontology for semantic web. In *Advances in Engineering, Science and Management (ICAESM), 2012 International Conference on* (pp. 400-405).
- [0201] [KhounSiavash & Baraani-Dastjerdi, 2010] KhounSiavash, E. & Baraani-Dastjerdi, A. (2010). Using the whole structure of ontology for semantic relatedness measurement. *SEKE*.
- [0202] [Leacock & Chodorow, 1998] Leacock, C. & Chodorow, M. (1998). Combining local context and WordNet similarity for word sense identification. In C. Fellbaum (Ed.), *WordNet: An electronic lexical database* (pp. 265-283). Massachusetts Institute of Technology.
- [0203] [Lee et al., 1993] Lee, J. H., Kim, M. H., & Lee, Y. J. (1993). Information retrieval based on conceptual distance in is-a hierarchies. *Journal of Documentation*, 49(2), 188-207.
- [0204] [Lee et al., 2008] Lee, W.-N., Shah, N., Sundlass, K., & Musen, M. (2008). Comparison of ontology-based semantic-similarity measures. *AMIA Annu. Symp. Proc.*, (pp. 384-388).
- [0205] [Lewis et al., 2004] Lewis, D. D., Yang, Y., Rose, T. G., & Li, F. (2004). RCV1: A new benchmark collection for text categorization research. *J. Mach. Learn. Res.*, 5, 361-397.
- [0206] [Lin, 1998] Lin, D. (1998). An information-theoretic definition of similarity. In *Proceedings of the 15th International Conference on Machine Learning*, volume 98 (pp. 296-304). Madison, Wis.
- [0207] [Lingling & Junzhong, 2012] Lingling, L. & Junzhong, J. (2012). A new model of information content based on concept's topology for measuring semantic similarity in WordNet1. *International Journal of Grid and Distributed Computing*, 5 (3), 81-94.
- [0208] [Lord et al., 2003] Lord, P. W., Stevens, R. D., Brass, A., & Goble, C. A. (2003). Semantic similarity measures as tools for exploring the gene ontology. *Pac. Symp. Biocomput.*, (pp. 601-612).
- [0209] [Machhour & Kassou, 2013] Machhour, H. & Kassou, I. (2013). Ontology integration approaches and its impact on text categorization. *International Journal of Data Mining & Knowledge Management Process (IJDKP)*, 3, 31-42.
- [0210] [McKeown et al., 1999] McKeown, K. R., Klavans, J. L., Hatzivassiloglou, V., Barzilay, R., & Eskin, E. (1999). Towards multidocument summarization by reformulation: Progress and prospects. In *Proceedings of the national conference on Artificial intelligence* (pp. 453-460). John Wiley & Sons Ltd.
- [0211] [Meng et al., 2012] Meng, L., Gu, J., & Zhou, Z. (2012). A review of information content metric for semantic similarity. In *Advances on Digital Television and Wireless Multimedia Communications, Communications in Computer and Information Science* (pp. 299-306). Springer, Berlin Heidelberg.
- [0212] [Meng et al., 2005] Meng, W., Xiaorong, W., & Chao, X. (2005). An approach to concept obtained text summarization. In *Communications and Information Technology, 2005. ISCIT 2005. IEEE International Symposium on*, volume 2 (pp. 1337-1340).
- [0213] [Metzler, 2007] Metzler, D. A. (2007, September). *Beyond bags of words: effectively modeling dependence and features in Information Retrieval* (PhD diss.). University of Massachusetts Amherst.
- [0214] [Mihalcea & Tarau, 2004] Mihalcea, R. & Tarau, P. (2004). TextRank: Bringing order into texts. In *Proceedings of EMNLP*, volume 4: Barcelona, Spain.

- [0215] [Miller, 1995] Miller, G. (1995). WordNet: A lexical database for English. *Commun. ACM*, 38 (11), 39-41.
- [0216] [Mustafa et al., 2008] Mustafa, J., Khan, S., & Latif, K. (2008). Ontology based semantic information retrieval. In *Intelligent Systems, 2008. IS.08. 4th International IEEE Conference*, volume 3 (pp. 22-14-22-19). Varna: IEEE.
- [0217] [Navigli, 2009] Navigli, R. (2009). Word sense disambiguation: A survey. *ACM Computing Surveys*, 41 (2), 10.
- [0218] [Orum & Joslyn, 2009] Orum, C. & Joslyn, C. A. (2009). Valuations and metrics on partially ordered sets. <http://arxiv.org/abs/0903.2679>
- [0219] [Pesquita et al., 2009] Pesquita, C., Faria, D., Falcao, A. O., Lord, P., & Couto, F. M. (2009). Semantic similarity in biomedical ontologies. *PLoS Computational Biology*, 5(7), e1000443.
- [0220] [Pierce, 1991] Pierce, B. C. (1991). *Basic Category Theory for Computer Science*. Cambridge, USA: Massachusetts Institute of Technology.
- [0221] [Pirró & Euzenat, 2010] Pirró, G. & Euzenat, J. (2010). A feature and 5 information theoretic framework for semantic similarity and relatedness. In *The Semantic Web Conference (ISWC. 2010), Lecture Notes in Computer Science* (pp. 615-630). Springer Berlin Heidelberg.
- [0222] [Pirró & Seco, 2008] Pirró, G. & Seco, N. (2008). Design, implementation and evaluation of a new semantic similarity metric combining features and intrinsic information content. In R. Meersman & Z. Tari (Eds.), *On the Move to Meaningful Internet Systems: OTM 2008*, volume 5332 of *Lecture Notes in Computer Science* (pp. 1271-1288).: Springer Berlin Heidelberg.
- [0223] [Pirró, 2009] Pirró, G. (2009). A semantic similarity metric combining features and intrinsic information content. *Data & Knowledge Engineering*, 68(11), 1289-1308.
- [0224] [Rada et al., 1989] Rada, R., Mili, H., Bicknell, E., & Blettner, M. (1989). Development and application of a metric on semantic nets. *IEEE Trans. Syst. Man Cybern.*, 19(1), 17-30.
- [0225] [Resnik, 1995] Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the IJCAI* (pp. 448-453).
- [0226] [Salton et al., 1975] Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Commun ACM*, 18(11), 613-620.
- [0227] [Sánchez & Batet, 2012] Sánchez, D., & Batet, M. (2012). A new model to compute the information content of concepts from taxonomic knowledge. *International Journal on Semantic Web and Information Systems*, 8(2), 34-50.
- [0228] [Sánchez et al., 2011] Sánchez, D., Batet, M., & Isern, D. (2011). Ontology-based information content computation. *Internat. J. Uncertain. Fuzziness Knowledge-Based Systems*, 24(2), 297-303.
- [0229] [Sánchez et al., 2012] Sánchez, D., Batet, M., Isern, D., & Valls, A. (2012). Ontology-based semantic similarity: A new feature-based approach. *Expert Syst. Appl.*, 39, 7718-7728.
- [0230] [Saruladha et al., 2011] Saruladha, K., Aghila, G., & others (2011). Information content based semantic similarity for cross ontological concepts. *International Journal of Engineering Science & Technology*, 3(6), 5132-5140.
- [0231] [Saruladha et al., 2010] Saruladha, K., Aghila, G., & Raj, S. (2010). A survey of semantic similarity methods for ontology based information retrieval. In *Machine Learning and Computing (ICMLC), 2010 Second International Conference on* (pp. 297-301). IEEE.
- [0232] [Saruladha et al., 2012] Saruladha, K., Aghila, G., & Raj, S. (2012). Semantic similarity measures for information retrieval systems using ontology. In *Second International Conference on Machine Learning and Computing (ICMLC), 2010* (pp. 297-301).: IEEE.
- [0233] [Sebastiani, 2002] Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys (CSUR)*, 34(1), 1-47.
- [0234] [Seco et al., 2004] Seco, N., Veale, T., & Hayes, J. (2004). An intrinsic information content metric for semantic similarity in WordNet. In R. López de Mántaras & L. Saitta (Eds.), *Proceedings of the 16th European Conference on Artificial Intelligence (ECAI)*, volume 16 (pp. 1089-1094). Valencia, Spain: IOS Press.
- [0235] [Siddharthan et al., 2004] Siddharthan, A., Nenkova, A., & McKeown, K. (2004). Syntactic simplification for improving content selection in multi-document summarization. In *Proceedings of the 20th international conference on Computational Linguistics* (pp. 896).: Association for Computational Linguistics.
- [0236] [Taieb et al., 2012] Taieb, M. A. H., Ben Aouicha, M., Tmar, M., & Ben Hamadou, A. (2012). Wikipedia category graph and new intrinsic information content metric for word semantic relatedness measuring. In *Data and Knowledge Engineering, Lecture Notes in Computer Science* (pp. 128-140). Springer Berlin Heidelberg.
- [0237] [Turney & Pantel, 2010] Turney, P. D. & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *J. Artif. Intell. Res.*, 37, 141-188.
- [0238] [Vallet et al., 2005] Vallet, D., Fernández, M., & Castells, P. (2005). An ontology-based information retrieval model. In *The Semantic Web: Research and Applications 2nd European Semantic Web Conference (ESWC 2005)* (pp. 455-470). Heraklion, Crete, Greece: Springer.
- [0239] [Vanderwende et al., 2004] Vanderwende, L., Banko, M., & Menezes, A. (2004). Event-centric summary generation. Working notes of DUC.
- [0240] [Widdows, 2004] Widdows, D. (2004). *Geometry and meaning*. CSLI publications Stanford.
- [0241] [Wolf & Gibson, 2004] Wolf, F. & Gibson, E. (2004). Paragraph-, word-, and coherence-based approaches to sentence ranking: A comparison of algorithm and human performance. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics* (pp. 383). Association for Computational Linguistics.
- [0242] [Wu et al., 2011] Wu, J., Ilyas, I., & Weddell, G. (2011). A study of ontology-based query expansion. Technical Report CS-2011-04, University of Waterloo.
- [0243] [Wu & Palmer, 1994] Wu, Z. & Palmer, M. (1994). Verbs semantics and lexical selection. In *Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics, ACL '94* (pp. 133-138). Stroudsburg, Pa., USA: Association for Computational Linguistics.
- [0244] [Xu & Shi, 2012] Xu, Q. & Shi, W. (2012). A comparison of semantic similarity models in evaluating concept similarity. In *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences (ISPRS.2012)*, volume XXXIX-B2, (pp. 173-178). Melbourne, Australia.

[0245] [Zhou et al., 2008] Zhou, Z., Wang, Y., & Gu, J. (2008). A new model of information content for semantic similarity in WordNet. In Future Generation Communication and Networking Symposia, 2008. FGCNS.08. Second International Conference on, volume 3 (pp. 85-89). IEEE.

SUMMARY OF THE INVENTION

[0246] The main object of the present invention is to address the problems, drawbacks and limitations of the current ontology-based information retrieval (IR) models reported in the literature, which have been previously described.

[0247] It is an object of the present invention to disclose a method and system to build a novel ontology-based information retrieval (IR) model for the indexing and retrieval of semantically annotated information units, that we call Intrinsic Ontological Spaces. Moreover, the present invention also comprises a novel ontology-based semantic distance called weighted Jiang-Conrath distance that overcomes some drawbacks of the classical Jiang-Conrath semantic distance in the context of the main problem addressed by this invention.

[0248] The purpose of the present method is to provide an ontology-based IR model for the indexing and retrieval of semantically annotated data, such as text documents, web pages, or any other sort of information units that can be represented by semantic annotations within a base ontology.

[0249] The main idea of the disclosed method and system is to build a structure-preserving embedding of a populated ontology into a metric space, while its intrinsic geometry is preserved. The proposed approach bridges the gap of modeling inconsistencies in current methods, achieving a better ranking, precision and recall measures for any indexing and retrieval system based on it.

[0250] The proposed IR model unifies the representation of the classes and individuals of the ontology in a same semantic space, while their intrinsic semantic structure relations are preserved. The text documents, or any other sort of semantically annotated units, are represented by sets of weighted-mentions to individuals and classes within the base ontology, while the queries are represented as sets of mentions to individuals and classes considered as sets of subsumed concepts and individuals. The representation space is a metric space defined by an extension of the novel weighted Jiang-Conrath distance between concepts, whose purpose is to define the distance among weighted mentions to individuals and classes, and a weighting scheme to represent documents and queries.

[0251] The intrinsic geometry of any ontology is defined by three algebraic structures: (1) the order relation of the taxonomy, (2) the intrinsic semantic distance among the classes and individuals, and (3) the set inclusion for the individuals and subsumed classes of the ontology.

[0252] The method proposed in this invention comprises the following elements:

[0253] (1) the definition of the semantic representation space as the universal set of weighted-mentions to individuals and classes within the populated base ontology, space that we call Intrinsic Ontological Spaces;

[0254] (2) an embedding method to inject semantically annotated data, or information units, in the representation space of the model;

[0255] (3) an embedding method to inject semantically annotated queries in the semantic representation space of the model;

[0256] (4) a semantic weighting method that combines statistical and semantic information to represent the semantic annotations associated to the indexed information units in the semantic representation space;

[0257] (5) a novel ontology-based semantic distance among concepts (classes) and instances of concepts (individuals) within a populated base ontology, that we call weighted Jiang-Conrath distance;

[0258] (6) a novel ontology-based ranking method for the retrieval and sorting of the indexed units retrieved by the system;

[0259] (7) a pre-processing step for computing all the parameters and data structures to enable the indexing and searching operations of the search engine of the system;

[0260] (8) a retrieval method to get a ranked collection of indexed units related to an input query; and

[0261] (9) an indexing and storing method to insert new data into the search and indexing system.

[0262] The representation space is a metric space defined by a distance function that is an extension of the novel weighted Jiang-Conrath distance disclosed in the present invention. The purpose of this novel ontology-based distance is to integrate the individuals in the representation, defining a metric on any sort of ontology, and to allow the definition of a semantic weighting scheme to represent documents and queries, while it overcomes the drawbacks of the standard Jiang-Conrath distance on taxonomies of concepts.

[0263] The weighted Jiang-Conrath distance is defined as the shortest weighted-path metric among concepts within the base ontology, according to the edge weights defined as the IC value of the joint probability between every child concept and its parent concept. The weighted Jiang-Conrath distance of the present invention is a well-defined metric on any sort of taxonomy. By contrast, the Jiang-Conrath distance is only a metric on tree-like ontologies [Orum & Joslyn, 2009]. The proposed model uses a novel intrinsic edge-based IC-based method, which will be later discussed in detail, to compute the IC-based edge weights used by the novel distance disclosed herein; however, we could use any other method to compute the edge weights, without changing the core of the invention herein proposed.

[0264] The IC-values depend only on the structure of the ontology; thus, these can be computed a priori during the set up process of the search engine. On the other hand, the computation of the shortest weighted-path among concepts needs to be done through any Djikstra-type algorithm, with the inconvenient that it could be very expensive for large base ontologies. For this reason, a pre-processing step is defined to compute a priori the IC-values and all the pair-wise distances among concepts of the base.

[0265] The semantic ranking method proposed in the present invention is based on a distance function between document and queries, which is defined by the Hausdorff distance among subsets of a metric space, according to the semantic metric of the representation space. The proposed model is a well-defined metric space on any sort of base ontology. By other hand, the use of the Hausdorff distance guarantees that the model is a well-defined metric space on the space of all annotated information units (documents) within any sort of base ontology.

[0266] The IR model herein proposed is the first one to use an ontology-based semantic distance and the Hausdorff distance as ranking method, unifying the representation of weighted-mentions to classes and individuals in a same met-

ric space. Moreover, in the case of a tree-like base ontology, the representation space also verifies the structure of a hierarchical Voronoi diagram [Gold & Angel, 2006], where every parent concept geometrically subsumes its descendant concepts. The present model transforms the logic hierarchy of the base ontology into a geometric hierarchy according to the semantic metric of the model.

[0267] The proposed model unifies the representation of the classes and individuals of the ontology in a same semantic metric space, while their intrinsic semantic relations are preserved. The documents, or other information units, are represented by sets of weighted-mentions to individuals and classes in the ontology, while the queries are represented as sets of mentions to individuals and classes considered as sets of subsumed concepts and individuals.

[0268] The present model avoids the use of vector spaces to rank documents; instead, a document (information unit) is defined as a collection of weighted mentions to classes and individuals, and the ranking method for documents is built using the metric of the space and the Hausdorff distance among subsets. The mentions to classes (concepts) within a user query are mapped to subsets of the representation space, while the mentions to classes in the documents are managed as weighted mentions to distinguished individuals of the parent class.

[0269] The main feature of the proposed model is that the embedding of the information units in the semantic representation space preserves the three main structure relations of the ontology, defined as follows: (1) the intrinsic semantic distance (metric structure) among classes, (2) the taxonomic relations (topologic/order structure), and (3) the set inclusion relations (set structure). We call these three structure relations as the intrinsic ontological structure.

[0270] The proposed model builds a natural equivalence between the information units and its embedding in the representation space, following the notion of natural morphism in the theory of categories [Eilenberg1945-am]. The proposed model tries to capture and to save all the semantic information provided by the ontology, avoiding any information lost in the embedding process. The input for our representation space is a populated ontology with semantic annotations of any sort of information unit; it means that the model assumes the existence of a semantic annotation module whose aim is to search the references to classes and entities of the ontology.

[0271] The Intrinsic Ontological Spaces model allows ranking documents, or any other semantically annotated data, using a semantic distance function derived from the ontology model, which improves the ranking quality and the precision and recall measures of the current methods, while it solves the inconsistencies in the current models.

[0272] The proposed solution has some theoretical and practical advantages over current methods. In particular:

[0273] First, the proposed IR model removes some inconsistencies in previous models, such as the orthogonality property and the cardinality mismatch, the lack of a ranking method based in an intrinsic semantic, and the lack of a semantic weighting method. The removal of these inconsistencies contributes to get an improved semantic representation model, whose main consequence is the improvement of the ranking quality and the precision and measures for any application based on the novel IR model.

[0274] Second, all the logic components of the IR model are ontology-based. It includes the retrieval process, the

weighting schema, the ranking and the definition of the representation space itself. Every element of the IR model is directly derived from the structure relations encoded by the base ontology used for the indexing of the data.

[0275] Third, the proposed IR model allows the integration of many geometry-based algorithms and theoretical results with potential benefits for the model. For example, we can integrate well known geometry-based space search methods to find nearby documents [Brin, 1995], enabling the extension of the model to large scale document collections as the web, or large text repositories in government agencies and private companies.

[0276] Fourth, the ranking and weighting computation model that is herein proposed can be estimated on-the-fly without any training phase, because the novel weighting method does not use an inverse frequency table.

[0277] Fifth, the factorization of the weights for the mentions in static (normalized frequency of the mentions) and dynamic (IC-value per concept) factors, allows updating the input parameters of the model (IC-values) while the index form of the indexed units is preserved. It means that the ontology could be dynamically updated in different ways (merge, concept insertion, etc.), without making changes to the indexed units, if the parent classes for the weighted-mentions are still in the ontology. For example, if a set of new classes is added to the base ontology, the system only needs to run the pre-processing step to get the new set of distances among concepts (classes); then, any new query answer will be computed using the novel semantic relations in the base ontology.

[0278] Sixth, like other known methods, the present model also merges the retrieval and ranking of documents in a same step, removing in this way the necessity to use SPARQL or any other query language to retrieve the documents to be ranked, as well as any other semantic retrieval method as it is proposed in [Mustafa et al., 2008]. The query is represented as any another document, and it is used to search the full set of indexed documents, eliminating the first retrieval step of many of the adapted VSM models.

[0279] Throughout the present discussion a language inspired in geometric notions is used, with the purpose to enlighten some analogies and relations between the conceptual spaces and its geometric images, expecting that it allows using all these well established and powerful theories. The present invention is related to some results in a novel research trend called geometry-based semantic models, whose first reference is defined by the preliminary ideas discussed in [Widdows, 2004] and partially developed in [Clarke, 2007].

[0280] As discussed above, the present invention is a novel contribution to the family of ontology-based IR models, and to the family of ontology-based semantic distances. Lastly, the Intrinsic Ontological Spaces can also be interpreted as an extreme way of ontology-based query expansion, problem recently revised in [Wu et al., 2011], where all the admissible query expansions are already integrated in the representation space itself, avoiding the query expansion problem. In this last direction, in [Saruladha et al., 2012] the authors studied how the semantic similarity/distance measures could be used to expand the user's query through the use of semantically related word in an ontology, problem that is avoided in the present invention.

[0281] In accordance with one aspect of the present invention there is provided a computer-implemented method for retrieving semantically relevant information units from a col-

lection of semantically annotated indexed information units in response to a query. The method comprises:

[0282] receiving, by a computer system, a semantically annotated query, the semantically annotated query including a set of semantic annotations to individuals or classes within a determined populated base ontology;

[0283] embedding, by the computer system, the semantically annotated query in a semantic representation space of an ontology-based IR model that uses a metric space for the representation of the indexed information units, the semantically annotated query being embedded as a set of weighted-mentions to individuals or classes within the populated base ontology;

[0284] obtaining, by the computer system, the representation in the semantic representation space for every indexed information unit of the collection;

[0285] computing, by the computer system, the Hausdorff distance between the space representation of the query and the space representation of all the indexed information units of the collection;

[0286] retrieving and ranking, by the computer system, the relevant information units based on the computed Hausdorff distance.

[0287] In a preferred embodiment the retrieving method may further comprise receiving, by the computer system, an input query in natural language; and converting, by the computer system, the input query into the semantically annotated query with regard to the populated base ontology.

[0288] A mention to a class in the semantically annotated query is preferably considered as a reference to a full class, the mentioned full class subsuming all the descendant individuals and classes within the populated base ontology.

[0289] The step of computing the Hausdorff distance preferably comprises:

[0290] retrieving information-content values of the populated base ontology;

[0291] retrieving pre-computed pair-wise semantic distances between all the classes within the base ontology;

[0292] computing all pair-wise semantic distances between the weighted-mentions of the query and the weighted-mentions of the indexed information units of the collection, using the information-content values and the pre-computed pair-wise semantic distances according to the metric of the representation space.

[0293] In a preferred embodiment the information-content values are retrieved by accessing a populated ontology repository.

[0294] The information-content values of the populated base ontology preferably include edge-based IC weights for each ontology edge, computed using an intrinsic edge-based IC-computation method.

[0295] The information-content values of the populated base ontology may include IC values for each ontology node computed using an intrinsic node-based IC-computation method.

[0296] The pre-computed pair-wise semantic distances are preferably retrieved by accessing a file.

[0297] In a preferred embodiment the pair-wise semantic distances between two ontology nodes are computed as the shortest weighted-path of the populated base ontology considering all possible paths between said two nodes, and that the edges of the taxonomy can be traversed in any direction.

[0298] The computation of pair-wise semantic distances between two ontology nodes preferably includes the compu-

tation of a weighted distance value which is the sum of the edge weights for all the edges of the populated base ontology along the shortest path joining said two ontology nodes.

[0299] In a preferred embodiment, the edge weight between two adjacent nodes of the populated base ontology is the information-content value of the joint probability between said two adjacent nodes.

[0300] The information-content value of the joint probability between two adjacent nodes is preferably computed as the negative binary logarithm of the joint probability.

[0301] In another preferred embodiment, the edge weight between two adjacent nodes of the populated base ontology is the information-content value of the child node minus the information-content value of the parent node.

[0302] The information units may be text documents, web pages, sentences, multimedia objects or any sort of data that can be represented as a collection of classes or individuals within an ontology.

[0303] The ontological representation space defined by the ontology-based IR model satisfies the following structure-preserving axioms: order invariance, metric invariance and inclusion invariance.

[0304] In accordance with a further aspect of the present invention there is provided a semantic search system for retrieving semantically relevant information units from a collection of semantically annotated indexed information units in response to a query, the semantic search system comprising a processor and a memory coupled with and readable by the processor and storing a set of instructions which, when executed by the processor, causes the processor to:

[0305] receive a semantically annotated query, the semantically annotated query including a set of semantic annotations to individuals or classes within a determined populated base ontology;

[0306] embed the semantically annotated query in a semantic representation space of an ontology-based IR model that uses a metric space for the representation of the indexed information units, the semantically annotated query being embedded as a set of weighted-mentions to individuals or classes within the populated base ontology;

[0307] obtain the representation in the semantic representation space for every indexed information unit of the collection;

[0308] compute the Hausdorff distance between the space representation of the query and the space representation of all the indexed information units of the collection;

[0309] retrieve and rank the relevant information units based on the computed Hausdorff distance.

[0310] The processor may be further configured to receive an input query in natural language; and convert the input query into the semantically annotated query with regard to the populated base ontology.

[0311] In accordance with yet a further aspect of the present invention there is provided a computer-readable memory for retrieving semantically relevant information units from a collection of semantically annotated indexed information units in response to a query, the computer-readable memory comprising a set of instructions stored therein which, when executed by a processor, causes the processor to:

[0312] receive a semantically annotated query, the semantically annotated query including a set of semantic annotations to individuals or classes within a determined populated base ontology;

[0313] embed the semantically annotated query in a semantic representation space of an ontology-based IR model that uses a metric space for the representation of the indexed information units, the semantically annotated query being embedded as a set of weighted-mentions to individuals or classes within the populated base ontology;

[0314] obtain the representation in the semantic representation space for every indexed information unit of the collection;

[0315] compute the Hausdorff distance between the space representation of the query and the space representation of all the indexed information units of the collection;

retrieve and rank the relevant information units based on the computed Hausdorff distance.

[0316] In accordance with yet a further aspect of the present invention there is provided a computer-implemented method for indexing semantically annotated information units into a search system based on an ontology-based information retrieval model, the method comprising:

[0317] receiving, by a computer system, a semantically annotated information unit, the semantically annotated information unit including a set of semantic annotations to individuals or classes within a determined populated base ontology and the frequency of said semantic annotations within the information unit;

[0318] embedding, by the computer system, the semantically annotated information unit in a semantic representation space of an ontology-based IR model that uses a metric space for the representation of the information units, the semantically annotated information unit being embedded as a set of weighted-mentions to individuals or classes within the populated base ontology;

[0319] storing the set of weighted-mentions of the indexed information unit in an information units repository;

[0320] storing the semantic annotations of the information unit in a populated ontology repository.

[0321] The indexing method may further comprise receiving, by the computer system, the information unit in natural language; and converting, by the computer system, the information unit into the semantically annotated information unit with regard to the populated base ontology.

[0322] The step of storing the semantic annotations of the information unit in a populated ontology repository preferably comprises:

[0323] inserting the new mentions to individuals and classes in the populated base ontology;

[0324] annotating the mentions to classes and individuals with inverted indexes to the indexed information units;

[0325] updating the information-content values for the registered individuals.

[0326] The information-content values for the registered individuals may be node-based IC values computed using an intrinsic node-based IC-computation method.

[0327] In a preferred embodiment the information-content values for the registered individuals are edge-based IC values computed using an intrinsic edge-based IC-computation method.

[0328] The edge-based IC values are preferably computed according to the updated joint probability between the individuals and their parent concepts.

[0329] In accordance with a further aspect of the present invention there is provided a system for indexing semantically annotated information units into a search system based on an ontology-based information retrieval model, the index-

ing system comprising a processor and a memory coupled with and readable by the processor and storing a set of instructions which, when executed by the processor, causes the processor to:

[0330] receive a semantically annotated information unit, the semantically annotated information unit including a set of semantic annotations to individuals or classes within a determined populated base ontology and the frequency of said semantic annotations within the information unit;

[0331] embed the semantically annotated information unit in a semantic representation space of an ontology-based IR model that uses a metric space for the representation of the information units, the semantically annotated information unit being embedded as a set of weighted-mentions to individuals or classes within the populated base ontology;

[0332] store the set of weighted-mentions of the indexed information unit in an information units repository;

[0333] store the semantic annotations of the information unit in a populated ontology repository.

[0334] In accordance with yet a further aspect of the present invention there is provided a computer-readable memory for indexing semantically annotated information units into a search system based on an ontology-based information retrieval model, the computer-readable memory comprising a set of instructions stored therein which, when executed by a processor, causes the processor to:

[0335] receive a semantically annotated information unit, the semantically annotated information unit including a set of semantic annotations to individuals or classes within a determined populated base ontology and the frequency of said semantic annotations within the information unit;

[0336] embed the semantically annotated information unit in a semantic representation space of an ontology-based IR model that uses a metric space for the representation of the information units, the semantically annotated information unit being embedded as a set of weighted-mentions to individuals or classes within the populated base ontology;

[0337] store the set of weighted-mentions of the indexed information unit in an information units repository;

[0338] store the semantic annotations of the information unit in a populated ontology repository.

BRIEF DESCRIPTION OF THE DRAWINGS

[0339] Embodiments of the present invention with now be described, by way of example, with reference to the accompanying drawings in which:

[0340] FIG. 1 shows the basic architecture of a multilingual or cross-language information retrieval (MLIR/CLIR) system as an application of the standard VSM IR models.

[0341] FIG. 2 shows a summary table with a comparative analysis of the ontology-based IR models used in the state of the art and in the present invention.

[0342] FIG. 3 represents a comparative analysis of semantic distances and similarities used in the state of the art and in the present invention.

[0343] FIG. 4 shows an example of a tree-like ontology, which is a partial sub-graph of WordNet around the "arm-chair" concept.

[0344] FIG. 5 shows an example of an algebraic lattice structure.

[0345] FIG. 6 shows a tree-like ontology with the edge weights defined as the difference of Information Content (IC) values between its extreme nodes.

[0346] FIG. 7 shows a taxonomy with upper semi-lattice structure.

[0347] FIG. 8 shows a taxonomy with the structure of a general partially ordered set (poset structure).

[0348] FIGS. 9A and 9B show a summary of the elements defined by the IR model proposed in the present invention.

[0349] FIG. 10 shows the geometric interpretation of the Hausdorff distance among subsets of a same metric space.

[0350] FIG. 11 shows the semantic representation of an input document into the metric space defined by the IR model proposed in the present invention.

[0351] FIG. 12 shows a representation about how the intrinsic order structure of a taxonomy is transformed in a geometric structure into the semantic representation space, which reproduces geometrically the intrinsic order structure of the taxonomy.

[0352] FIG. 13 shows an ontology example used for describing the proposed model.

[0353] FIG. 14A shows a flowchart of a preferred embodiment of the pre-processing step of the proposed IR model. FIG. 14B shows a flowchart of another embodiment of the pre-processing step of the proposed IR model.

[0354] FIG. 15 shows a flowchart describing the method to compute a set of ranked information unit as answer to an input query provided by any user.

[0355] FIG. 16 shows a flowchart describing the indexing method for any novel information unit to store in the system ontology-based repository.

PREFERRED EMBODIMENT OF THE INVENTION

[0356] The Intrinsic Ontological Spaces are a sort of semantic representation spaces for populated ontologies, which are based in a metric space derived from a novel ontology-based distance introduced in the present invention, and called weighted Jiang-Conrath distance. These spaces represent semantically annotated information units, such as text documents, in a metric space compound by weighted-elements. The semantic annotations encode mentions to individuals or classes within a base ontology provided by the user. The base ontology is populated with new added data through an indexing and storing process.

[0357] The representation space is a metric space endowed with a hierarchical structure, which represents the classes (concepts) and individuals (instances) in a base ontology within the same space, while their intrinsic structure relations are preserved.

[0358] The intrinsic geometry of the ontology is defined by three structures: (1) the semantic distance (metric structure) among classes, (2) the taxonomic relations (graph/order structure), and (3) the set inclusion relations (set structure).

[0359] The main goal of the present invention is to design a semantic representation of the ontology space that preserves these structures. The representation of the ontology is defined by a consistent metrization of the ontology space, integrating classes and individuals in a same representation, while their intrinsic structure relations are preserved.

[0360] The indexing method and IR model proposed in the present invention comprise the following components:

[0361] (1) the definition of the semantic representation space as the universal set of weighted-mentions to individuals and classes within the populated base ontology, space that we call Intrinsic Ontology Spaces;

[0362] (2) a novel ontology-based semantic distance among concepts (classes) and instances of concepts (individuals) within a populated base ontology, that we call weighted Jiang-Conrath distance;

[0363] (3) a semantic weighting method that combines statistical and semantic information to represent the semantic annotations associated to the indexed information units in the semantic representation space;

[0364] (4) an embedding method to injects semantically annotated data, or information units, in the representation space of the model;

[0365] (5) an embedding method to inject semantically annotated queries in the semantic representation space of the model;

[0366] (6) a novel ontology-based ranking method for the retrieval and sorting of the indexed units retrieved by the system;

[0367] (7) a pre-processing step for computing all the parameters and data structures to enable the indexing and searching operations of the search engine of the system;

[0368] (8) a retrieval method to get a ranked collection of indexed units related to an input query; and

[0369] (9) an indexing and storing method to insert new data into the search and indexing system.

[0370] We herein explain the proposed method and model by describing every component enumerated above, using the drawings. The summary of algebraic objects and formulas are summarized in the summary table represented in FIGS. 9A and 9B.

Definition of the Representation Space

[0371] Hereinafter, uppercase letters are used to denote sets, and lowercase letters are used for elements of a set. By the pair $T=(C, \leq_C)$, see row 1 in FIG. 9A, we denote a partially ordered set with a root node, which defines a taxonomy denoted by T, where $C=\{C_i\}$ is a non-empty finite set of concepts (classes) and \leq_C denotes the order relation on the set C.

[0372] If we provide the set of concepts C with any ontology-based metric $d_C:C \times C \rightarrow \mathbb{R}$, we get the metric space (C, d_C) , structure that we call a conceptual metric space, such as is defined below. Hence, A conceptual metric space is a pair (C, d_C) defined by a non-empty set of concepts $C=\{C_i\}$ endowed with a metric $d_C:C \times C \rightarrow \mathbb{R}$ on it.

[0373] We can extend the taxonomy T with a family of sets $I_C=\{I_{C_i}\}$ (see row 2 in FIG. 9A) of instances (individuals) to concepts in C, to obtain the pair (CUI_{C_i}, \leq_C) , called a populated base ontology (see row 3 in FIG. 9A). A populated ontology is the tuple $O=(C \cup I_C, \leq_C)$, where (C, \leq_C) is a taxonomy and $I_C=\{I_{C_i}\}$ is a family of sets of instances to concepts in C.

[0374] Endowing the populated ontology with any ontology-based metric $d_C:C \times C \rightarrow \mathbb{R}$, we get the tuple $\mathcal{O}=(CUI_{C_i}, \leq_C, d_C)$, structure that we call a metric ontology, as defined in row 4 of FIG. 9A. Hence, a metric ontology is a tuple $\mathcal{O}=(CUI_{C_i}, \leq_C, d_C)$, where the tuple (CUI_{C_i}, \leq_C) is a populated ontology, where there is also a metric space defined by the distance function $d_C:C \times C \rightarrow \mathbb{R}$. The finite set $I_C=\{I_{C_i} | \forall C_i \in C\}$ is a family of sets whose elements are instances of every concept $C_i \in C$. The family of sets of instances verifies the set inclusion relation with regard to its associated concepts, such that it verifies $\forall C_i \in C \rightarrow I_{C_i} \subset C_i$.

[0375] The goal of the IR model proposed in this invention is to build a semantic representation for any semantically

annotated corpus, whose semantic annotation is based in a populated ontology (CUI_C, \leq_C). The method only considers the “is-a” relations within the ontology, discarding any other sort of relations. The “is-a” relation refers to the explicit inheritance relation between two concepts, by example, one motorbike is a two-wheel vehicle, thus, the last one concept subsumes the first one. The input to our IR model is a populated taxonomic ontology.

[0376] We define an ontology representation space, as the pair (X, d_X) , with $X = \text{CUI}_C \times [0, 1] = \mathbb{R}$, where $d_X: X \times X \rightarrow \mathbb{R}$ is a metric on the space of weighted-mentions to individuals and classes within a base metric ontology. The elements of the representation space X are weighted references to instances of classes (individuals), or weighted references to concepts (classes). The ontology representation space, shown in row 7 of FIG. 9A, is thus considered a metric space (X, d_X) , where $X = \text{CUI}_C \times [0, 1] = \mathbb{R}$ is a space of weighted-mentions to individuals and classes on a base metric ontology $\mathcal{O} = (\text{CUI}_C, \leq_C, d_C)$, and $d_X: X \times X \rightarrow \mathbb{R}$ is a metric on X .

[0377] In the present model, any information unit is represented as a set of weighted references to classes or individuals. The inputs to the model are a collection of information units annotated with classes and individuals in the populated ontology. The information units could be documents, or other information source that admits the same representation. The information units must have been semantically annotated with the frequency of typed entities (individuals) or classes within the base ontology, as shown in row 5 of FIG. 9A. The weighting method used in the model consists in the unit normalization of the frequencies of every typed reference in the document (TF-weighting). As shown in row 6 of FIG. 9A, any annotated information unit δ_k is a set of tuples $\delta_k = \{(\tau_j, f_j^k) \in \text{CUI}_C \times \mathbb{N} \mid j \in J(k)\}$ where τ_j denotes the j -th reference to a class or typed individual in the ontology, f_j^k the frequency of the concept or instance τ_j in the information unit δ_k , and $J(k)$ is a set of indexes of the individuals or classes cited in the information unit δ_k .

[0378] The ontology metric space is defined by the metric $d_X: X \times X \rightarrow \mathbb{R}$, which is based in an extension of the novel ontology-based distance called weighted Jiang-Conrath given by equation (3) below.

[0379] Once the main elements of the present IR model have been introduced, we define the first principles or axioms satisfied by the model. The intrinsic ontology embedding and Intrinsic Ontology Spaces are defined in row 8 of FIG. 9A. Given a populated ontology (CUI_C, \leq_C) , its associated metric ontology $\mathcal{O} = (\text{CUI}_C, \leq_C, d_C)$, a space of frequency-annotated information units $D = \text{CUI}_C \times \mathbb{N}$, and a metric space (X, d_X) . A function pair $\phi_f: D \rightarrow X$ and $\phi_c: C \rightarrow X$ is called an intrinsic ontological embedding, and the metric space (X, d_X) an Intrinsic Ontological Space, if the following axioms are satisfied:

[0380] (1) Order (subsumption) invariance:

$$C_1 \leq_C C_2 \Rightarrow \phi_c(C_1) \subset \phi_c(C_2), \forall (C_1, C_2) \in C \times C$$

[0381] (2) Metric invariance:

$$d_C(C_1, C_2) = d_H(\phi_c(C_1), \phi_c(C_2)), \forall (C_1, C_2) \in C \times C$$

[0382] (3) Inclusion invariance:

$$\phi_f(\tau) \subset \phi_c(C_i), \forall \tau \in I_{C_i} \quad a)$$

$$\phi_f(\tau) \subset \phi_c(C_j), \forall \tau \in I_{C_i}, \forall C_j \mid C_i \leq_C C_j \quad b)$$

[0383] In this way, the Intrinsic Ontology Spaces are defined as an ontological representation space which verifies the structure-preserving axioms enumerated above. Note that

in the proposed model, every individual is considered as a child node of its parent concept, with its own IC-value. By other hand, every full class within the ontology is mapped as a set to the representation space through the function ϕ_c .

[0384] The first axiom states that the mapping ϕ_c preserves the subset (order) relation among concepts in the representation space, while the taxonomy order is transformed in a space subset relation according to the metric d_X . The second axiom establishes the natural equivalence (morphism) between the input ontology-based metric d_C and the metric on subsets of the representation space, given by the Hausdorff distance d_H on the space of subsets of X . It means that the distance between concepts in the ontology is equal to the distance among its images. Finally, the third axiom establishes that the image of every individual (instance) of class in the representation space, must be included in the image set of its parent class and all the ancestor classes within the ontology. The last axiom (3.b) can be deduced from the axioms (1) and (3.a).

[0385] The algebraic objects involved in the definition of the Intrinsic Ontology Spaces model are summarized in FIGS. 9A and 9B.

[0386] The proposed method in this invention to build the IR model called Intrinsic Ontology Spaces comprises: (1) the definition of the collection of algebraic objects and functions summarized in FIGS. 9A and 9B, (2) the pre-processing step described in FIGS. 14A and 14B, (3) the search and retrieval method described in FIG. 15, and (4) the indexing method of information units described in FIG. 16. These mathematical objects are defined as follows.

[0387] Individuals Embedding.

[0388] The function ϕ_f defined in row 9 of FIG. 9A defines the embedding for isolated mentions to individuals within the populated ontology in the metric space X . If one information unit contains only one mention to an individual or class within the populated ontology, the individual is assigned the static weight 1, without taking into account its frequency inside the indexed unit.

[0389] Whole Class Embedding.

[0390] In the case of queries, we consider that any mention to a class (concept) within the populated ontology is a reference to the full class, it means that the query selects all the classes (concepts) and individuals (instances) subsumed by the mentioned class. Precisely, it is the meaning of the embedding function ϕ_c given by the definition of row 10 in FIG. 9A. The image of any class C_i is simply the image in the representation space X of every subsumed class or individuals within the ontology.

[0391] Information Unit Embedding.

[0392] The function ϕ , defined in the row 11 of FIG. 9B, defines the embedding and static weighting method used to represent the information units (documents) in the representation space of the proposed IR model. The input to the function is a set of frequency-based weighted mentions to individuals and classes inside the information units to be indexed. The function computes a set of static weights for each mention through the normalization of the frequencies in the unit to be indexed, or represented in the space (X, d_X) of the model. The function generates a set of tuples representing the weighted-mentions as indexing representation for the unit, such as is described in FIG. 16.

[0393] Semantic Weighting.

[0394] Because an information unit (document) is a collection of frequency-based weighted mentions, doing some alge-

bra, we can conclude that the information content (IC-value) of the information unit is the sum of the IC-value of every mention multiplied by its number of occurrences in the information unit. The last observation allows us to define the normalized IC-value for each indexed information unit, as shown in row **12** of FIG. 9B. The normalized IC-value is a combination of two factors: (1) the static weights, which can be interpreted as the statistical fingerprint of the information unit, and (2) the IC-value of the mentioned node which is a semantic descriptor, updated dynamically every time a new information unit is indexed by the IR model. The normalized IC-value endorses the use of static normalized term-frequency (TF) weights as index form for the indexed information units, as shown in row **11** of FIG. 9B.

[0395] Novel Weighted Jiang-Conrath Distance.

[0396] The drawbacks of the standard Jiang-Conrath distance were already mentioned. To overcome these limitations and extending ontology-based distances to individuals, we herein introduce a novel ontology-based distance that we call weighted Jiang-Conrath distance, denoted by d_{wJC} , which is defined in row **13** of FIG. 9B. The name of the distance recalls that it is an extension and generalization of the standard Jiang-Conrath distance to any sort of taxonomy. Given any populated ontology $O=(CUI_C, \leq c)$, we define a weighted and unoriented graph $G=(V, E, w)$, with a positive real-valued function on each edge $w: E \rightarrow \mathbb{R}$. Every vertex $v \in V$ represents an element in the set CUI_C , and the edges are defined by $E = \{(ci, cj) \in CUI_C \times CUI_C \mid ci \neq cj, \wedge cj = LA(ci)\}$, where LA is the lowest ancestor. The edge-based weight function $w(e_{ij})$ is defined as the IC-value of the joint probability of observation for a subsumed child concept ci given its parent concept cj , as it is defined in equation (3). The weighted Jiang-Conrath distance d_{wJC} is defined as the shortest weighted-path among two taxonomy nodes $a, b \in CUI_C$, as it is shown in equation (3) given below.

[0397] The evaluation of d_{wJC} requires the implementation of any shortest path algorithm to compute the shortest path among nodes of the ontology, such as the revised ones in [Ahuja et al., 1990]. Equation (3) represents the novel weighted Jiang-Conrath distance.

$$d_{wJC}: C \times C \rightarrow \mathbb{R} \quad (3)$$

$$d_{wJC}(a, b) = \min_{x \in P(a, b)} \left\{ \sum_{e_{ij} \in x} w(e_{ij}) \right\}$$

$$w: E \rightarrow \mathbb{R}^+$$

$$w(e_{ij}) = IC(P(c_i | c_j)) = -\log_2 P(c_i | c_j)$$

$$P(c_i | c_j) \in [0, 1] \rightarrow w(e_{ij}) \geq 0, \forall e_{ij} \in E$$

[0398] In equation (3) the distance value is the sum of weights for all the edges along the shortest path joining the ontology nodes “a” and “b”, which can be concepts or individuals, considering the taxonomy as an unoriented graph. $P(a, b)$ denotes all possible edge paths joining the nodes “a” and “b” within the base ontology. The edge weights are defined by the IC-value of the joint probability $P(ci|cj)$ between any child concept ci and its parent concept cj . Note that edge weights are always positive, because the joint probability takes values in the range $[0, 1]$. The definition of the edge weights is valid on any sort of taxonomy, and if the

taxonomy is tree-like, the weights induce that d_{wJC} mimics the classical Jiang-Conrath distance.

[0399] Computation of the Edge Weights.

[0400] The edge weights $w(e_{ij})$ can be computed in two different forms as follows: (1) using an intrinsic IC-computation method, or (2) using an intrinsic computation method for the edge-based joint probabilities, the latter being the preferred method. It is now explained how to compute the edge weights using these two different approaches:

[0401] First, if any intrinsic IC-computation method is used to compute the IC values in the nodes, such as the method of [Zhou et al., 2008] defined by the equation (4) or other equivalent method proposed in [Seco et al., 2004], then the edge weights are defined by the equation (5). In equation (4) $hypo(c)$ is the number of child concepts (hyponyms) of the concept c , k is a constant in the range $[0, 1]$ used to balance the contribution of every term (Zhou et al. suggests $k=0.5$), $node_{max}$ is the total number of concepts in the taxonomy, $deep(c)$ is the depth of the concept in the taxonomy (the distance measured as the number of nodes from the concept to the root node), and $deep_{max}$ is the maximum depth value for any concept in the taxonomy.

$$IC(c) = -\log(p(c)) = k \left(1 - \frac{\log(hypo(c) + 1)}{\log(node_{max})} \right) + (1 - k) \frac{\log(deep(c))}{\log(deep_{max})} \quad (4)$$

$$w(e_{ij}) = IC(c_i) - IC(c_j) \quad (5)$$

[0402] Second, a direct method to estimate intrinsically the joint probabilities $P(ci|cj)$ for each edge can be used, fulfilling the intrinsic property that the sum of all the joint probabilities for the subsumed concepts of any parent concepts must be 1, as it is expressed by the equation (6). In this second case, new methods are herein provided to compute intrinsically the joint probabilities, and thus, the edge weights can be computed using the formula $w(e_{ij}) = IC(P(ci|cj))$ in equation (3), wherein the joint probabilities $P(ci|cj)$ are defined, given by the equations (7), (8) and (9) below, being the equation (8) the preferred method. This method for the computation of the edge weights admits any other intrinsic estimation method for the edge-based joint probabilities that could be developed in the future. The computation of the IC-based edge weights using the equation (7) is called IC-JointProbUniform, while the method using the equation (8) is called IC-JointProbHypo, and when the equation (9) is used the method is called IC-JointProbLeaves.

$$\sum_{\forall |c_j = LA(c_i)} P(c_i | c_j) = 1 \quad (6)$$

$$P(c_i | c_j) = \frac{1}{|children(c_j)|} \rightarrow w(e_{ij}) = \log_2 |children(c_j)| \quad (7)$$

$$P(c_i | c_j) = \quad (8)$$

$$\frac{hypo(c_i) + 1}{\sum_{\forall |c_j = LA(c_i)} (hypo(c_i) + 1)} \rightarrow w(e_{ij}) = -\log_2 \left(\frac{hypo(c_i) + 1}{\sum_{\forall |c_j = LA(c_i)} (hypo(c_i) + 1)} \right)$$

-continued

$$P(c_i | c_j) = \frac{\text{leaves}(c_i) + 1}{\sum_{v|c_j=LA(c_i)} (\text{leaves}(c_i) + 1)} \rightarrow w(e_{ij}) = -\log_2 \left(\frac{\text{leaves}(c_i) + 1}{\sum_{v|c_j=LA(c_i)} (\text{leaves}(c_i) + 1)} \right) \quad (9)$$

[0403] The equation (7) represents a uniform probability distribution for each direct child (1-degree descendant) concept, and $|\text{children}(ci)|$ is the number of direct child concepts. By other hand, the (8) and (9) equations are two weighting schemes based in the proportion of subsumed concepts or leaves for each child concept (node). The edge-based joint probability $P(c_i|c_j)$ could be intrinsically computed by any other method, but it doesn't change the definition of the ontology-based semantic distance described herein.

[0404] In equation (8) the function $\text{hypo}(ci)$ defines the number of subsumed concepts by the concept ci without inclusive it, while the function $\text{leaves}(ci)$ in equation (9) returns the number of leaf concepts (no descendants) subsumed by the concept (ci) without inclusive it.

[0405] The intrinsic IC-computation methods associated to the equation (7), (8) and (9) could be easily adapted to be used in combination with any other ontology-based IC semantic measure in two steps: (1) the computation of the probability for each concept in the taxonomy, and (2) the computation of the IC values for each concept given by the negative binary logarithm of the probabilities values of each concept. One method for the first step is to use a total ordering traversal of the taxonomy starting on the root node, then to compute the probability for each concept traversing the ordered list of concepts. Here we describe in detail the previous method. First, the probability computation method must build a total ordering of the concept in the taxonomy, which is defined by an ordered list of concepts, such that every concept is in a subsequent position to every one of its parent concepts. Second, the method assigns a value of "1" to the probability of the first concept in the ordered list (root concept), then, the method traverses the concept nodes according to the total ordering built before, and it computes the probability of every child concept as the sum over each parent of the product of the parent probability by estimated intrinsic joint probability $P(\text{child}|\text{parent})$.

[0406] Extending the Distance Among Concepts to Individuals.

[0407] In the proposed model, every individual is considered as a child node of its parent concept, with its own IC-value. It means that the distance from an individual to any other ontology node is simply the accumulated weighted-path from the individual to the extreme node, which can be a class or other individual.

[0408] Static and Dynamic Nodes of the Base Ontology.

[0409] The populated ontology of the IR model contains static and dynamic nodes. The static nodes correspond to the classes within the base ontology, while the dynamic nodes correspond to the individuals, which are inserted as semantic annotations for the indexed info units. The proposed model considers two different types of individuals: (1) mentions to instances of concepts (individuals), or (2) mentions to whole classes using any name, synonym or morphological variant of the classes, for example, the word "hospital" in the user query

"hospitals in Washington", refers to all the instances subsumed by the concept "hospital" that are registered within the populated ontology. Due to its static nature, we can compute a priori all the pair-wise distance values among classes of the base ontology, such as it is described in FIGS. 14A and 14B. The last process is called the pre-processing step, and it allows the efficient computation in run-time of the distance among weighted-mentions in the representation space.

[0410] Metric of the Intrinsic Ontology Space.

[0411] In row 14 of FIG. 9B we can appreciate the distance function d_x which defines the metric for the representation space of the proposed IR model. The metric d_x defines the distance between weighted-mentions to individuals or classes in the representation space (X, d_x) . The d_x function is given by the sum of distances from every weighted-mention to its lowest ancestor (LA) within the base ontology, plus the pre-computed weighted Jiang-Conrath distance among the LA concepts of both mentions. The present model supports multiple inheritance for any individuals, which means that any individual can belong to more than one class; thus, the closed form for the distance function d_x is the shortest weighted-path considering all possible lowest ancestors of any individual, such as is defined by the formula shown in row 14 of FIG. 9B. The metric of the space given by d_x is defined by two terms according to the difference among the lowest ancestor classes of every pair of mentions. Given two weighted-mentions (x, w_x) and (y, w_y) , if x is equal to y , then the distance value is given by the absolute value of the binary $\log(w_x/w_y)$; otherwise, the distance d_x is defined by the minimum value (shortest path) over the Cartesian product $LA(x) \times LA(y)$, such as shown in the formula of row 14 in FIG. 9B.

[0412] Intrinsic Ontological Spaces (IOS).

[0413] The representation space of the IR model proposed in the present invention is called Intrinsic Ontological Spaces. As it is shown in row 15 of FIG. 9B, the IOS space is simply the metric space (X, d_x) according to the function d_x defined above.

[0414] Hausdorff Distance Between Subsets of a Metric Space.

[0415] The Hausdorff distance is well-known in the literature about metric spaces, whose definition is included in row 16 of FIG. 9B. The symbol $P(X)$ denotes the power set of the set X , compounded by all the subsets of X . We use the Hausdorff distance to define the ontology-based ranking method d_D of the IR model proposed in this invention, as shown in row 17 of FIG. 9B. FIG. 10 shows the geometric interpretation of the Hausdorff distance among subsets of a same metric space. The Hausdorff distance is defined as the largest distance value among all the pair-wise minimum distance value from the elements from one subset X to the opposite subset Y . For each subset, the Hausdorff distance selects the element with the largest distance value to the opposite subset; then, the Hausdorff distance is defined as the maximum of these two values. For example, the left bottom circle is centred on the element of Y with the largest distance value to any element of set X . On the other hand, the right bottom circle is centred on the element of X with the largest distance value to the any element of the set Y . Therefore, the Hausdorff distance will be the largest distance value between the distance values associated to these elements.

[0416] Ontology-Based Ranking.

[0417] In row 17 of FIG. 9B we define the ranking function as the distance function d_D among information units. Any input query is embedded through the process described in

FIG. 15. Once the query has been injected in the representation space, the system measures its distance to every indexed information unit using the function d_D , which is simply their Hausdorff distance according to the metric d_X . The ranking is ontology-based because it is based in the Hausdorff distance, which is derived from d_X , and the last one is also derived from an ontology-based semantic distance, the novel weighted Jiang-Conrath distance.

[0418] Distance Between any Weighted-Mention and a Full Class.

[0419] We already explained above that the mention to any class (concept) within an input query is considered as a reference to the full class. It means that any mentioned full class must subsume all its descendant nodes within the ontology. The key idea to speed-up the computation of the Hausdorff in this case is to realize that we can use the definition of the full class embedding ϕ_C , given by the expression in row 10 of FIG. 9A. Following this observation, we arrive to definition of the distance d_{IC} (individual-full class), shown in row 18 of FIG. 9B. This function is used in step 1522 of the search and retrieval process described in FIG. 15. Note that a full class is selecting the region of the representation space that includes all its descendant nodes. The function d_{IC} works like a geometry-based logic operator which avoids the necessity of executing a preliminary formal query on the populated ontology, with the aim of retrieving the indexed units semantically that are related to the input query.

[0420] FIG. 11 shows, as an example, a document embedding in the Intrinsic Ontological Space. The mentions to concepts (classes) and instances of concepts (individuals) within the document are weighted and injected in the metric space holding its relations to its parent concepts (classes). Since a metric space is a free-coordinate space, we do not have a direct image of the elements, as in the case of an Euclidean representation. By contrast, we can only appreciate the structure of the space through the distance relations induced by the metric. In FIG. 11, the acronym NE stands for Named Entity, being it the common name assigned in the scope of IR to the instances of a defined concept.

[0421] FIG. 12 shows the intrinsic embedding on any populated base ontology in the representation space. Every concept (class) of the taxonomy is mapped to a subset of the representation space which subsumes the images by the embedding of all its individuals and derived concepts. If the taxonomy is tree-like, then we get a hierarchical Voronoi structure [Good & Angel, 2006], where each cell is defined by the site associated to the distinguished element for each class. If one concept "E" derives from two parent concepts "C" and "D", then its image in the representation space is a region in the intersection of both parent regions.

[0422] FIG. 13 shows an ontology example in the bioengineering domain which is used as base ontology in the rest of the figures to clarify the operation of the proposed method. The ontology defines some types of documents to be indexed, as well as some types of entities, pathologies and sort of agents to be recognized within the indexed papers. The ontology defines an object model to organize the content of the indexed documents and for supporting the search and retrieval process.

[0423] The pre-processing, search and indexing methods will now be described in detail. The pre-processing of the IR model is described in FIGS. 14A and 14B; the search and retrieval method is described in FIG. 15; and the indexing method is described in FIG. 16.

[0424] Pre-Processing Method.

[0425] FIGS. 14A and 14B show two different preferred embodiments of a pre-processing method (1400a, 1400b) for any indexing system based in the proposed IR model, whose main goal is computing all the static parameters required for the operation (loading and execution) of the proposed IR model. These parameters are, in the particular embodiment of FIG. 14A, the edge-based Information Content (IC) weights for the ontology edges and all pair-wise semantic distances among the classes of the ontology. In the embodiment of FIG. 14B these static parameters are the IC values for each ontology node and all pair-wise semantic distances among the classes of the ontology. The all pair-wise semantic distances correspond to the values of the function d_{wJC} in row 13 of FIG. 9B. In a preferred embodiment the edge-based IC-weights of the embodiment of FIG. 14A are computed using the method previously disclosed, whose formula is given by equation (8) above. In the embodiment of FIG. 14B the IC values are computed using an intrinsic node-based IC-computation method (for instance, using the formula of equation (4)).

[0426] In step 1404a of the embodiment of FIG. 14A and step 1404b of the embodiment of FIG. 14B the system administrator 1402 provides a base ontology, for instance in OWL XML format file or other equivalent format. In step 1406a/1406b the system reads the input file and loads the base ontology in the computer memory. In step 1408a the system computes the joint probabilities values for each edge (FIG. 14A) using the preferred method, for instance defined by the equation (8), while in step 1408b the system computes the IC values for each class node (FIG. 14B) within the ontology using the equation (4). In step 1410a the system assigns the computed joint probabilities to each edge (FIG. 14A), while in step 1410b the system assigns de IC values for each node (FIG. 14B) of the base ontology. In step 1412a of FIG. 14A the system computes all the IC-based edge weights values for the edges of the ontology as the negative binary logarithm of the joint probabilities for each edge, such as is defined by the edge-valued function $w(e_i)$ in the equation (3), generating the weighted-graph 1414a, which is stored in the computer memory, and a XML file 1416a containing the data representation for the weighted-graph associated to the base ontology. In step 1412b the system computes all the edge weights values for the edges of the ontology (IC value of the child minus IC value of the parent), generating two different outputs: the weighted-graph 1414b stored in the computer memory, and a XML file 1416b containing the data representation for the weighted-graph associated to the base ontology. In step 1418a/1418b the system computes all the pair-wise distances among concepts of the weighted-graph, using the novel weighted Jiang-Conrath distance denoted by d_{wJC} , which is defined as the shortest path between two concepts, as shown in row 13 of FIG. 9B. The shortest path can be computed using any known variant of the Dijkstra algorithm, or any other method proposed in the literature. For the computation of all pair-wise distance among concepts within the weighted-graph, the system can use the representation of the weighted-graph 1414a/1414b stored in the computer memory and any shortest path implemented algorithm, or it can invoke any external program or library using the information in the XML file 1416b. In step 1420a/1420b, the system represents the distance matrix 1422a/1422b in the computer memory whenever it is possible, or uses any file-storing method to save this information. Finally, in step

1424a/1424b, the system generates the final pre-processing files as follows: an OWL XML file **1426a/1426b** with the base ontology plus the IC-values for each edge (FIG. 14A) or node (FIG. 14B), and a XML file **1428a/1428b** containing the pair-wise semantic distances for the base ontology.

[0427] To summarize, the pre-processing goal is to compute all the data needed for the loading and execution of the proposed IR mode as follows:

[0428] (1) the computation **1408a/1408b** of all the edge-based joint probabilities or node-based IC-values, using for instance the method defined by equation (7) or equation (8), and the equation (4), where $\text{hypo}(c)$ is the number of subsumed child or descendants concepts (hyponyms) of the concept c , whereas $|\text{children}(c)|$ in equation (7) refers only to the number of direct child concepts; and

[0429] (2) the computation **1418a/1418b** of all the pair-wise distance among the classes within the base ontology provided by the users, which can be computed by any graph-based shortest path algorithm as the revised ones in [Abuja et al., 1990] or any variant of the well-known Dijkstra's algorithm.

[0430] Search and Retrieval Method.

[0431] FIG. 15 shows the flowchart of the search and retrieval method **1500** to be used by any search system based in the proposed IR model. The user **1502** first provides an input query **1504** in text format, or other symbolic representation, which is fed into the semantic search system **1505** based in Intrinsic Ontology Spaces. Then, in step **1506** the semantic search system **1505** uses any automatic semantic annotator, out of the scope of the present invention, to convert the input query **1504** into a set of semantic annotations (ontology-based semantic annotation of the query **1508** or, in short, annotated query **1508**) to individuals or classes within the populated base ontology of the system. In step **1510** the annotated query **1508** is embedded in the Intrinsic Ontological Space defined in row **15** of FIG. 9B. In the step **1510**, id represents the unique identifier of any class or individual in the populated ontology, and the integer field "1" encodes a whole mention to any concept or individual within the populated base ontology. In step **1512** the retrieval and ranking of the relevant indexed information units (documents), with respect to the input query, is performed. The retrieval and ranking is based in the Hausdorff distance among the query and the indexed units. In step **1514**, the semantic search system **1505** computes the ontology-based ranking function d_D defined by row **17** in FIG. 9B. The ranking d_D is defined by the Hausdorff distance (d_H) between the representation of the query (step **1510**) and the representation of every indexed document. The representation of the queries and documents is defined by a set of weighted-mentions to classes and individuals. The computation of the Hausdorff distance requires the computation of all pair-wise distances among the weighted-mentions of the queries and the weighted-mentions of the indexed documents, according to the metric of the representation space, given by d_X in row **14** of FIG. 9B. In step **1516** the semantic search system **1505** gets the space representation for every indexed unit included in the information units repository **1518**, where every indexed unit is represented by a set of weighted-mentions to classes or individuals. For the computation of pair-wise distances, we distinguish two cases as follows:

[0432] For mentions to individuals in a query, in step **1520** the distance between weighted-mention to individuals and classes is computed using the formula for d_X in row **14** of FIG. 9B.

[0433] For mentions to (full) classes in a query, in step **1522** the distance between weighted-mentions and full classes is computed using the formula d_{IC} in row **18** of FIG. 9B. For example, the distance between the whole mention to the class drug (Drug, 1) in step **1510** and any weighted-mention to class or individual within the populated ontology, is computed using the formula d_{IC} in row **18** of FIG. 9B.

[0434] In step **1524** the semantic search system **1505** gets the edge-based IC-values (weights) for the weighted-mentions to individuals and classes (which are used in the formula for d_X), accessing a populated ontology repository **1526**, which includes the pre-computed OWL XML file **1426a** with the base ontology and the IC-values (for the embodiment of FIG. 14B the semantic search system **1505** gets the IC values). In step **1528** the semantic search system **1505** uses the distances matrix loaded in memory, by accessing the pre-computed XML file **1428a** containing the pair-wise semantic distances for the base ontology. Finally, the semantic search system **1505** performs the distance-based ranking **1530** and outputs a set of ranked information units **1532** semantically related to the query.

[0435] To sum up, the main goal of the method represented in FIG. 15 is the retrieving of a set of ranked documents as answer to any input query provided by the users. The documents are ranked according to their semantic similarity with the semantic representation of any input query. The steps **1506**, **1508** and **1510** in the flowchart have as main purpose to convert the input query into a semantic information unit in the semantic space defined by the proposed IR model. One specific feature of the present approach is that the mentions to classes in the query are interpreted as references to the whole class; thus, the classes are embedded in the representation space as geometric regions which subsume all their descendant nodes within the populated ontology of the indexing system. This can be appreciated in step **1508**, where the mention to the "Drug" concept has been defined as a whole class "class: Drug". The step **1512** includes the Hausdorff distance **1514** to compare the query with all the indexed units, and the distance-based ranking **1530** to sort the results according to the measured distance in the representation space. The Hausdorff distance uses two different functions to measure the distance among the query and any indexed unit: the metric of the representation space to compute the distance among weighted-mention to individuals or classes (step **1520**); and the distance among any weighted-mention to individual or class and any whole class (step **1522**). The distance function in step **1520** is defined by the function in row **14** of FIG. 9B, whereas the distance function in step **1522** is the defined in row **18** of FIG. 9B.

[0436] Indexing Method of Semantically Annotated Information Units.

[0437] FIG. 16 shows the method for indexing **1600** any additional information unit (e.g. a new document) into a search system based on the proposed model, and then storing it in the system ontology-based repository. First, in step **1604** a user **1602** provides a new information unit, such as a text document **1606**, to be registered by the indexing system **1608**. In step **1610** the same automatic semantic annotator used in the query process (in step **1506** of FIG. 15) is used to identify the mentions to individuals and classes which are present, or

able to be represented, in the populated base ontology. The semantic annotation step **1612** produces a set of semantic annotations plus its frequency within the input document. In another preferred embodiment these steps **1610** and **1612** are performed externally to the indexing system **1608**, such that the indexing system **1610** retrieves the semantic annotations and frequency directly from an automatic semantic annotator. In step **1614** the document is embedded (represented) in the Intrinsic Ontological Spaces as a set of normalized weighted-mentions to classes and individuals in the base ontology. The indexing step is split in two:

[0438] A first indexing step **1616** has as main aim the storing of the index form, defined by a set of static semantic weights (normalized frequency of the mentions), into the repository **1518** of the indexed information units.

[0439] A second indexing step **1618** has as main goal the storing of the semantic annotations of the input document in the populated ontology repository **1526**. The new mentions to individuals and classes are inserted in the populated ontology (sub-step **1620**). The new annotations are extended to keep the inverted index to the indexed units where it appears (sub-step **1622**). The edge-based IC-values (for the embodiment of FIG. **14A**) or node-based IC-values (for the embodiment of FIG. **14B**) for the registered individuals are updated (sub-step **1624**), in the first case according to the updated joint probability between the individuals and their parent concepts.

[0440] To summarize, in the example shown in FIG. **16** the user provides a research paper in the bioengineering field as input document **1606** to be indexed by the indexing system **1608**. An automatic semantic annotator, out of the scope of this invention, identifies mentions to classes and individuals according to the conceptual model defined by the base ontology. Then, the semantic annotator produces a frequency-based semantic annotation of the document (step **1612**). In step **1614** the indexing system **1608** computes the embedding of the document in the representation space, which consists in a collection of normalized weighted-mentions to the classes and individuals stored within the base ontology. The collection of tuples (node type, node identifier, static weight) defines a whole index for each unit indexed by the system. The indexing system **1608** stores **1616** the index form for the inserted document in the indexing repository **1518**, updates the populated ontology to store the new individuals (sub-step **1620**), to annotate the classes and individuals with back references to the indexed units (sub-step **1622**), and to update the joint probabilities and IC-values of the individuals (sub-step **1624**). It is important to note that the IC-values for the edges linking the classes are static, and their value has been already computed in the pre-processing step described in FIG. **14**.

We claim:

1. A computer-implemented method for retrieving semantically relevant information units from a collection of semantically annotated indexed information units in response to a query, the method comprising:

receiving, by a computer system, a semantically annotated query, the semantically annotated query including a set of semantic annotations to individuals or classes within a determined populated base ontology;

embedding, by the computer system, the semantically annotated query in a semantic representation space of an ontology-based IR model that uses a metric space for the representation of the indexed information units, the

semantically annotated query being embedded as a set of weighted-mentions to individuals or classes within the populated base ontology;

obtaining, by the computer system, the representation in the semantic representation space for every indexed information unit of the collection;

computing, by the computer system, the Hausdorff distance between the space representation of the query and the space representation of all the indexed information units of the collection;

retrieving and ranking, by the computer system, the relevant information units based on the computed Hausdorff distance.

2. The computer-implemented method of claim **1**, further comprising:

receiving, by the computer system, an input query in natural language;

converting, by the computer system, the input query into the semantically annotated query with regard to the populated base ontology.

3. The computer-implemented method of claim **1**, wherein a mention to a class in the semantically annotated query is considered as a reference to a full class, the mentioned full class subsuming all the descendant individuals and classes within the populated base ontology.

4. The computer-implemented method of claim **1**, wherein the step of computing the Hausdorff distance comprises:

retrieving information-content values of the populated base ontology;

retrieving pre-computed pair-wise semantic distances between all the classes within the base ontology;

computing all pair-wise semantic distances between the weighted-mentions of the query and the weighted-mentions of the indexed information units of the collection, using the information-content values and the pre-computed pair-wise semantic distances according to the metric of the representation space.

5. The computer-implemented method of claim **4**, wherein the information-content values are retrieved by accessing a populated ontology repository.

6. The computer-implemented method of claim **4**, wherein the information-content values of the populated base ontology include IC values for each ontology node computed using an intrinsic node-based IC-computation method.

7. The computer-implemented method of claim **4**, wherein the information-content values of the populated base ontology include edge-based IC weights for each ontology edge, computed using an intrinsic edge-based IC-computation method.

8. The computer-implemented method of claim **4**, wherein the pre-computed pair-wise semantic distances are retrieved by accessing a file.

9. The computer-implemented method of claim **4**, wherein the pair-wise semantic distances between two ontology nodes are computed as the shortest weighted-path of the populated base ontology considering all possible paths between said two nodes, and that the edges of the taxonomy can be traversed in any direction.

10. The computer-implemented method of claim **4**, wherein the computation of pair-wise semantic distances between two ontology nodes includes the computation of a weighted distance value which is the sum of the edge weights for all the edges of the populated base ontology along the shortest path joining said two ontology nodes.

11. The computer-implemented method of claim 10, wherein the edge weight between two adjacent nodes of the populated base ontology is the information-content value of the joint probability between said two adjacent nodes, wherein the sum of all the joint probabilities for the subsumed concepts of any parent concept is equal to 1.

12. The computer-implemented method of claim 11, wherein the information-content value of the joint probability between two adjacent nodes is computed as the negative binary logarithm of the joint probability.

13. The computer-implemented method of claim 10, wherein the edge weight between two adjacent nodes of the populated base ontology is the information-content value of the child node minus the information-content value of the parent node.

14. The computer-implemented method of claim 1, wherein the information units are text documents, web pages, sentences, multimedia objects or any sort of data that can be represented as a collection of classes or individuals within an ontology.

15. The computer-implemented method of claim 1, wherein the ontological representation space defined by the ontology-based IR model satisfies the following structure-preserving axioms: order invariance, metric invariance and inclusion invariance.

16. A semantic search system for retrieving semantically relevant information units from a collection of semantically annotated indexed information units in response to a query, the semantic search system comprising a processor and a memory coupled with and readable by the processor and storing a set of instructions which, when executed by the processor, causes the processor to:

receive a semantically annotated query, the semantically annotated query including a set of semantic annotations to individuals or classes within a determined populated base ontology;

embed the semantically annotated query in a semantic representation space of an ontology-based IR model that uses a metric space for the representation of the indexed information units, the semantically annotated query being embedded as a set of weighted-mentions to individuals or classes within the populated base ontology;

obtain the representation in the semantic representation space for every indexed information unit of the collection;

compute the Hausdorff distance between the space representation of the query and the space representation of all the indexed information units of the collection;

retrieve and rank the relevant information units based on the computed Hausdorff distance.

17. The system of claim 16, wherein the processor is further configured to:

receive an input query in natural language;

convert the input query into the semantically annotated query with regard to the populated base ontology.

18. The system of claim 15, wherein a mention to a class in the semantically annotated query is considered by the processor as a reference to a full class, the mentioned full class subsuming all the descendant individuals and classes within the populated base ontology.

19. The system of claim 16, wherein the computing of the Hausdorff distance comprises:

retrieving information-content values of the populated base ontology;

retrieving pre-computed pair-wise semantic distances between all the classes within the base ontology;

computing all pair-wise semantic distances between the weighted-mentions of the query and the weighted-mentions of the indexed information units of the collection, using the information-content values and the pre-computed pair-wise semantic distances according to the metric of the representation space.

20. The system of claim 19, wherein the information-content values are retrieved by accessing a populated ontology repository.

21. The system of claim 19, wherein the information-content values of the populated base ontology include IC values for each ontology node computed using an intrinsic node-based IC-computation method.

22. The system of claim 19, wherein the information-content values of the populated base ontology include edge-based IC weights for each ontology edge, computed using an intrinsic edge-based IC-computation method.

23. The system of claim 19, wherein the pre-computed pair-wise semantic distances are retrieved by accessing a file.

24. The system of claim 19, wherein the pair-wise semantic distances between two ontology nodes are computed as the shortest weighted-path of the populated base ontology considering all possible paths between said two nodes, and that the edges of the taxonomy can be traversed in any direction.

25. The system of claim 19, wherein the computation of pair-wise semantic distances between two ontology nodes includes the computation of a weighted distance value which is the sum of the edge weights for all the edges of the populated base ontology along the shortest path joining said two ontology nodes.

26. The system of claim 25, wherein the edge weight between two adjacent nodes of the populated base ontology is the information-content value of the joint probability between said two adjacent nodes, wherein the sum of all the joint probabilities for the subsumed concepts of any parent concept is equal to 1.

27. The system of claim 26, wherein the information-content value of the joint probability between two adjacent nodes is computed as the negative binary logarithm of the joint probability.

28. The system of claim 25, wherein the edge weight between two adjacent nodes of the populated base ontology is the information-content value of the child node minus the information-content value of the parent node.

29. The system of claim 16, wherein the information units are text documents, web pages, sentences, multimedia objects or any sort of data that can be represented as a collection of classes or individuals within an ontology.

30. The system of claim 16, wherein the ontological representation space defined by the ontology-based IR model satisfies the following structure-preserving axioms: order invariance, metric invariance and inclusion invariance.

31. A computer-readable memory for retrieving semantically relevant information units from a collection of semantically annotated indexed information units in response to a query, the computer-readable memory comprising a set of instructions stored therein which, when executed by a processor, causes the processor to:

receive a semantically annotated query, the semantically annotated query including a set of semantic annotations to individuals or classes within a determined populated base ontology;

embed the semantically annotated query in a semantic representation space of an ontology-based IR model that uses a metric space for the representation of the indexed information units, the semantically annotated query being embedded as a set of weighted-mentions to individuals or classes within the populated base ontology;
 obtain the representation in the semantic representation space for every indexed information unit of the collection;
 compute the Hausdorff distance between the space representation of the query and the space representation of all the indexed information units of the collection;
 retrieve and rank the relevant information units based on the computed Hausdorff distance.

32. A computer-implemented method for indexing semantically annotated information units into a search system based on an ontology-based information retrieval model, the method comprising:

receiving, by a computer system, a semantically annotated information unit, the semantically annotated information unit including a set of semantic annotations to individuals or classes within a determined populated base ontology and the frequency of said semantic annotations within the information unit;

embedding, by the computer system, the semantically annotated information unit in a semantic representation space of an ontology-based IR model that uses a metric space for the representation of the information units, the semantically annotated information unit being embedded as a set of weighted-mentions to individuals or classes within the populated base ontology;

storing the set of weighted-mentions of the indexed information unit in an information units repository;

storing the semantic annotations of the information unit in a populated ontology repository.

33. The computer-implemented method of claim **32**, further comprising:

receiving, by the computer system, the information unit in natural language;

converting, by the computer system, the information unit into the semantically annotated information unit with regard to the populated base ontology.

34. The computer-implemented method of claim **32**, wherein the step of storing the semantic annotations of the information unit in a populated ontology repository comprises:

inserting the new mentions to individuals and classes in the populated base ontology;

annotating the mentions to classes and individuals with inverted indexes to the indexed information units;

updating the information-content values for the registered individuals.

35. The computer-implemented method of claim **34**, wherein the information-content values for the registered individuals are node-based IC values computed using an intrinsic node-based IC-computation method.

36. The computer-implemented method of claim **34**, wherein the information-content values for the registered individuals are edge-based IC values computed using an intrinsic edge-based IC-computation method.

37. The computer-implemented method of claim **36**, wherein the edge-based IC values are computed according to the updated joint probability between the individuals and their parent concepts.

38. The computer-implemented method of claim **32**, wherein the ontological representation space defined by the ontology-based IR model satisfies the following structure-preserving axioms: order invariance, metric invariance and inclusion invariance.

39. A system for indexing semantically annotated information units into a search system based on an ontology-based information retrieval model, the indexing system comprising a processor and a memory coupled with and readable by the processor and storing a set of instructions which, when executed by the processor, causes the processor to:

receive a semantically annotated information unit, the semantically annotated information unit including a set of semantic annotations to individuals or classes within a determined populated base ontology and the frequency of said semantic annotations within the information unit;

embed the semantically annotated information unit in a semantic representation space of an ontology-based IR model that uses a metric space for the representation of the information units, the semantically annotated information unit being embedded as a set of weighted-mentions to individuals or classes within the populated base ontology;

store the set of weighted-mentions of the indexed information unit in an information units repository;

store the semantic annotations of the information unit in a populated ontology repository.

40. The system of claim **39**, wherein the processor is further configured to:

receive the information unit in natural language;

convert the information unit into the semantically annotated information unit with regard to the populated base ontology.

41. The system of claim **39**, wherein the storing the semantic annotations of the information unit in a populated ontology repository comprises:

inserting the new mentions to individuals and classes in the populated base ontology;

annotating the mentions to classes and individuals with inverted indexes to the indexed information units;

updating the information-content values for the registered individuals.

42. The system of claim **41**, wherein the information-content values for the registered individuals are node-based IC values computed using an intrinsic node-based IC-computation method.

43. The system of claim **41**, wherein the information-content values for the registered individuals are edge-based IC values computed using an intrinsic edge-based IC-computation method.

44. The system of claim **43**, wherein the edge-based IC values are computed according to the updated joint probability between the individuals and their parent concepts.

45. The system of claim **39**, wherein the ontological representation space defined by the ontology-based IR model satisfies the following structure-preserving axioms: order invariance, metric invariance and inclusion invariance.

46. A computer-readable memory for indexing semantically annotated information units into a search system based on an ontology-based information retrieval model, the computer-readable memory comprising a set of instructions stored therein which, when executed by a processor, causes the processor to:

receive a semantically annotated information unit, the semantically annotated information unit including a set of semantic annotations to individuals or classes within a determined populated base ontology and the frequency of said semantic annotations within the information unit;
embed the semantically annotated information unit in a semantic representation space of an ontology-based IR model that uses a metric space for the representation of the information units, the semantically annotated information unit being embedded as a set of weighted-mentions to individuals or classes within the populated base ontology;
store the set of weighted-mentions of the indexed information unit in an information units repository;
store the semantic annotations of the information unit in a populated ontology repository.

* * * * *

Part III

Software Libraries and Datasets

Chapter 12

HESML V1R2 Semantic Measure Library

This page intentionally left blank.

HESML V1R2 Java software library of ontology-based semantic similarity measures and information content models

Published: 21 Dec 2016 | **Version 2** | DOI: 10.17632/t87s78dg78.2

Contributor(s): [Juan J. Lastra-Díaz](#), [Ana Garcia-Serrano](#)

Description of this data

HESML V1R2 is the second release of the Half-Edge Semantic Measures Library (HESML) [1], which is a new, scalable and efficient Java software library of ontology-based semantic similarity measures and Information Content (IC) models based on WordNet.

HESML V1R2 implements most ontology-based semantic similarity measures and Information Content (IC) models based on WordNet reported in the literature. In addition, it provides a XML-based input file format in order to specify the execution of reproducible experiments on WordNet-based similarity, even with no software coding.

The V1R2 release significantly improves the performance of HESML V1R1. HESML is introduced and detailed in a companion reproducibility paper [1] of the methods and experiments introduced in [2,3,4].

The main features of HEMSL are as follows: (1) it is based on an efficient and linearly scalable representation for taxonomies called PosetHERep introduced in [1], (2) its performance exhibits a linear scalability as regards the size of the taxonomy, and (3) it does not use any caching strategy of vertex sets.

HESML V1R2 is freely distributed for any non-commercial purpose under a CC By-NC-SA-4.0 license, subject to the citing of the main HESML paper [1] as attribution requirement. On other hand, the commercial use of the similarity measures introduced in [2], as well as part of the intrinsic IC models introduced in [3] and [4], is protected by a patent application [5]. In addition, any user of HESML must fulfill other licensing terms described in [1] related to other resources distributed with the library, such as WordNet and a dataset of corpus-based IC models, among others.

References:

- [1] Lastra-Díaz, J. J., & García-Serrano, A. (2016). HESML: a scalable ontology-based semantic similarity measures library with a set of reproducible experiments and a replication dataset. To appear in *Information Systems Journal*.
- [2] Lastra-Díaz, J. J., & García-Serrano, A. (2015). A novel family of IC-based similarity measures with a detailed experimental survey on WordNet. *Engineering Applications of Artificial Intelligence Journal*, 46, 140–153.
- [3] Lastra-Díaz, J. J., & García-Serrano, A. (2015). A new family of information content models with an experimental survey on WordNet. *Knowledge-Based Systems*, 89, 509–526.
- [4] Lastra-Díaz, J. J., & García-Serrano, A. (2016). A refinement of the well-founded Information Content models with a very detailed experimental survey on WordNet. *Universidad Nacional de*

[5] Lastra Díaz, J. J., & García Serrano, A. (2016). System and method for the indexing and retrieval of semantically annotated data using an ontology-based information retrieval model. United States Patent and Trademark Office (USPTO) Application, US2016/0179945 A1.

Experiment data files

[Download all files \(1\)](#)

 HESML_Release_V1R2.zip

45 MB

HESML V1R2 Java source files

Steps to reproduce

HESML V1R2 is distributed as a Java class library (HESML-V1R2.jar) plus a test driver application (HESMLclient.jar), which have been developed using NetBeans 8.0.2 for Windows, although it has been also compiled and evaluated on Linux-based platforms using the corresponding NetBeans versions.

In order to compile HESML V1R2, you must follow the following steps:

- (1) Download the ZIP file above containing the full distribution of HESML V1R2..
- (2) Install Java 8, Java SE Dev Kit 8 and NetBeans 8.0.2 or higher in your workstation.
- (3) Launch NetBeans IDE and open the HESML and HESMLclient projects contained in the root folder. NetBeans automatically detects the presence of a nbproject subfolder with the project files.
- (4) Select HESML and HESMLclient projects in the project treeview respectively. Then, invoke the "Clean and Build project (Shift + F11)" command in order to compile both projects.

In order to remain up to date on new HESML versions, as well as asking for technical support, we invite the readers to subscribe to the HESML forum by sending an email to the following address:

hesml+subscribe@googlegroups.com

For more information, we refer the reader to the paper below:

Lastra-Díaz, J. J., & García-Serrano, A. (2016). HESML: a scalable ontology-based semantic similarity measures library with a set of reproducible experiments and a replication dataset. To appear in Information Systems.

Related links

[A novel family of IC-based similarity measures with a detailed experimental survey on WordNet](#)

article is related to this dataset

doi:10.1016/j.engappai.2015.09.006

[A new family of information content models with an experimental survey on WordNet](#)

article is related to this dataset

Chapter 13

HESML V1R1 Semantic Measure Library

This page intentionally left blank.

You are viewing a previous version of this dataset. Version 2 is the latest version of this dataset.

HESML V1R1 Java software library of ontology-based semantic similarity measures and information content models

Published: 7 Sep 2016 | **Version 1** | DOI: 10.17632/t87s78dg78.1

Contributor(s): [Juan J. Lastra-Díaz](#), [Ana Garcia-Serrano](#)

Description of this data

HESML V1R1 is a new Java software library called Half-Edge Semantic Measures Library (HESML), which implements most ontology-based semantic similarity measures and Information Content (IC) models based on WordNet reported in the literature.

HESML is introduced and detailed in the paper by Lastra-Díaz, J. J., & García-Serrano, A. (2016). HESML: a scalable ontology-based semantic similarity measures library with a set of reproducible experiments and a replication dataset. Information Systems.

HESML is motivated by several drawbacks in the current state-of-the-art software libraries, as well as the evaluation of the new methods introduced by the authors, together with the replication and evaluation of most previously reported methods.

HESML is based on a new and efficient poset representation, called PosetHERep, which is an adaptation of the half-edge data structure commonly used to represent discrete manifolds and planar graphs in computational geometry. HESML proposes a memory-efficient representation for taxonomies which linearly scales with the taxonomy size and provides an efficient implementation of a large set of topological queries and graph-based algorithms. Likewise, HESML provides an open framework to aid research into the area by providing a simpler and more efficient software architecture than the current software libraries.

Experiment data files

[Download all files \(1\)](#)

 HESML_Release_V1R1.zip

105 MB

HESML V1R1 Java source code

Steps to reproduce

- (1) Download the ZIP file above containing the full distribution of HESML V1R1.
- (2) Follow the instructions in the paper below:

Lastra-Díaz, J. J., & García-Serrano, A. (2016). HESML: a scalable ontology-based semantic similarity measures library with a set of reproducible experiments and a replication dataset. Information Systems.

Chapter 14

WNSimRep V1 dataset

This page intentionally left blank.

WNSimRep: a framework and replication dataset for ontology-based semantic similarity measures and information content models

Published: 8 Sep 2016 | **Version 1** | DOI: 10.17632/mpr2m8pycs.1

Contributor(s): [Juan J. Lastra-Díaz](#), [Ana Garcia-Serrano](#)

Description of this data

The WNSimRep v1 dataset is provided as supplementary material of the paper by Lastra-Díaz, J. J., & García-Serrano, A. (2016). HESML: a scalable ontology-based semantic similarity measures library with a set of reproducible experiments and a replication dataset. Information Systems.

In the aforementioned work, we introduce a scalable Java software library of ontology-based semantic similarity measures and IC models, called HESML, and a set of reproducible experiments on word similarity.

The WNSimRep v1 dataset is detailed in the enclosed file called "appendixB_WNSimRep_dataset_LastraGarcia_v1.pdf".

This work introduces a framework whose aim is to allow the exact replication of most intrinsic Information Content (IC) models and ontology-based similarity measures reported in the literature by using the publicly available accompanying dataset, called the WNSimRep v1 dataset. This work has been carried-out in the context of a large evaluation campaign of ontology-based semantic similarity measures and IC models on WordNet based on HESML. Our work is encouraged by the identification of several reproducibility problems in a series of recent experimental surveys carried-out by the authors, together with the lack of a framework and gold standard to assist in the replication of ontology-based similarity measures and IC models. To bridge this gap, we introduce herein a replication framework defined by three different types of data file: (a) node-based data files which contain an explicit representation of the WordNet taxonomy together with a specific IC model and a collection of node-based taxonomical features, (b) edge-based data files which contain a family of edge-valued IC models based on the conditional probability between child and parent concepts, and (c) synset-pair-based data files which contain the synset pairs of the Rubenstein-Goodenough word similarity benchmark, together with a collection of taxonomical features based on synset pairs and all the ontology-based similarity measures evaluated on them. The framework is implemented in the accompanying dataset which includes a collection of intrinsic and corpus-based IC models based on WordNet 3.0, enriched with a broad set of taxonomical features used by most intrinsic IC models and ontology-based similarity measures.

Experiment data files

[Download all files \(2\)](#)

WNSimRepV1.zip

109 MB

The WNSimRep v1 dataset file contains a collection of comma separated files (*.csv) which contain a rich set of taxonomical features and Information Content models based on WordNet 3.0 and the Rubenstein-Goodenough benchmark whose main aim is

 appendixB_WNSimRep_dataset_LastraGarcia.pdf

273 KB

This paper introduces a detailed description of the WNSimRep v1 dataset and a framework whose aim is to allow the exact replication of most intrinsic Information Content models and ontology-based similarity measures reported in the literature.

Steps to reproduce

The WNSimRep v1 dataset has been created by using the HESML library introduced in the main companion paper below:

[1] Lastra-Díaz, J. J., & García-Serrano, A. (2016). HESML: a scalable ontology-based semantic similarity measures library with a set of reproducible experiments and a replication dataset. Information Systems.

Thus, in order to reproduce the WNSimRep v1 dataset, you should follow the next steps:

- (1) Obtain a copy of the paper [1]
- (2) Follow the instructions in the section on reproducible experiments.

Related links

[A novel family of IC-based similarity measures with a detailed experimental survey on WordNet](#)

article is related to this dataset

doi:10.1016/j.engappai.2015.09.006

[A new family of information content models with an experimental survey on WordNet](#)

article is related to this dataset

doi:10.1016/j.knosys.2015.08.019

<http://e-spacio.uned.es/fez/view/bibliuned:DptoLSI-ETSI-Informes-Jlastra-refinement>

article is related to this dataset

<http://dx.doi.org/10.17632/t87s78dg78.1>

dataset is related to this dataset

<http://dx.doi.org/10.17632/65pxgskhz9.2>

Latest version

Version 1

2016-09-08

Published: 2016-09-08

DOI: 10.17632/mpr2m8pycs.1

Cite this dataset

Lastra-Díaz, Juan J.; Garcia-Serrano, Ana (2016), "WNSimRep: a framework and replication dataset for ontology-based semantic similarity measures and information content models", Mendeley Data, v1

<http://dx.doi.org/10.17632/mpr2m8pycs.1>

Institutions

National University of Distance Education

Categories

Similarity Measure, Ontological Models

Licence

CC BY NC 3.0

[Learn more](#)

[Report](#)

We care about your feedback

Help us to improve Mendeley Data by telling us what we can do better.

[Send feedback](#)

Mission

Archive Policy

Suggested file formats



ELSEVIER

[Copyright](#)

[Terms of Use](#)

[Privacy Policy](#)

Copyright © 2017 Mendeley Ltd. All rights reserved. Cookies are set by this site. To decline them or learn more, visit our [cookies page](#).

 **RELX** Group™

WNSimRep: a framework and replication dataset for ontology-based semantic similarity measures and information content models

Juan J. Lastra-Díaz Ana García-Serrano
(jlastra@invi.uned.es, agarcia@lsi.uned.es)

NLP and IR Research Group
E.T.S.I. Informática - UNED
Universidad Nacional de Educación a Distancia
C\ Juan del Rosal 16, 28040 Madrid (Spain)

July 25, 2016

Abstract

This paper introduces a framework whose aim is to allow the exact replication of most intrinsic Information Content (IC) models and ontology-based similarity measures reported in the literature by using the publicly available accompanying dataset, called the *WNSimRep v1* dataset. This work is a companion paper provided as supplementary material of Lastra-Díaz and García-Serrano (2016b). In this latter work, we introduce a scalable Java software library of ontology-based semantic similarity measures and IC models, called HESML, and a set of reproducible experiments on word similarity. This work has been carried-out in the context of a large evaluation campaign of ontology-based semantic similarity measures and IC models on WordNet based on HESML. Our work is encouraged by the identification of several reproducibility problems in a series of recent experimental surveys carried-out by the authors, together with the lack of a framework and gold standard to assist in the replication of ontology-based similarity measures and IC models. To bridge this gap, we introduce herein a replication framework defined by three different types of data file: (a) node-based data files which contain an explicit representation of the WordNet taxonomy together with a specific IC model and a collection of node-based taxonomical features, (b) edge-based data files which contain a family of edge-valued IC models based on the conditional probability between child and parent concepts, and (c) synset-pair-based data files which contain the synset pairs of the Rubenstein and Goodenough (1965) word similarity benchmark, together with a collection of taxonomical features based on synset pairs and all the ontology-based similarity measures evaluated on them. The framework is implemented in the accompanying dataset which includes a collection of intrinsic and corpus-based IC models based on WordNet 3.0, enriched with a broad set of taxonomical features used by most intrinsic IC models and ontology-based similarity measures.

Keywords: WNSimRep v1 dataset, intrinsic Information Content models, corpus-based Information Content models, ontology-based semantic similarity measures, IC-based similarity measures, replication similarity measures and IC models, WordNet-based similarity benchmarks.

1 Introduction

An ontology-based semantic similarity measure is a binary concept-valued function $sim : C \times C \rightarrow \mathbb{R}$ defined over a single-root taxonomy of concepts (C, \leq_C) which returns the degree of similarity between concepts as perceived by a human being. The current ontology-based semantic measures can be categorized into four families as follows: (1) edge-counting similarity measures, so called path-based measures, such as the pioneering work of Rada et al. (1989), whose core idea is the use of the length of the shortest path between concepts as an estimation of their degree of similarity; (2) IC-based similarity measures whose core idea is the use of an Information Content (IC) model, such as the pioneering work of Resnik (1995); (3) feature-based similarity mea-

asures, whose core idea is the use of set-theory operators between the feature sets of the concepts, such as the pioneering work of Tversky (1977), and more recently Sánchez et al. (2012), whose core idea is the use of the overlapping of ancestor sets as an estimation of the overlapping between the unknown feature sets of the concepts; (4) other similarity measures that cannot be directly categorized into any previous family, which are based on taxonomical features derived from set-theory operators Batet et al. (2011), or novel contributions of the hyponym set Hadj Taieb et al. (2014b).

Every IC-based similarity measure requires a complementary concept-valued function to be evaluated, which is called the Information Content (IC) model. Given a taxonomy of concepts defined by a triplet $C = ((C, \leq_C), \Gamma)$, where $\Gamma \in C$ is the supreme ele-

Reference	Definition of the non IC-based similarity measures
	$sim_{Rada}(c_1, c_2) = 1 - \frac{1}{2}d_{Rada}(c_1, c_2)$
Rada et al. (1989)	$d_{Rada}(c_1, c_2) = len(c_1, c_2) = \min_{\forall \alpha \in Paths_{(c_1, c_2)}} \left\{ \sum_{e_{ij} \in \alpha} 1 \right\}$
Wu and Palmer (1994)	$sim_{W\&P}(c_1, c_2) = \frac{2 \times depth(LCA(c_1, c_2))}{len(c_1, LCA(c_1, c_2)) + len(c_2, LCA(c_1, c_2)) + 2 \times depth(LCA(c_1, c_2))}$
Leacock and Chodorow (1998)	$sim_{L\&C}(c_1, c_2) = -\log \left(\frac{1 + len(c_1, c_2)}{2 \times maxdepth} \right)$
Li et al. (2003)	$sim_{Li_s3}(c_1, c_2) = e^{-\alpha^* len(c_1, c_2)}, \quad \alpha^* = 0.25$
Li et al. (2003)	$sim_{Li_s4}(c_1, c_2) = e^{-\alpha^* len(c_1, c_2)} \times \frac{e^{\beta^* d} - e^{-\beta^* d}}{e^{\beta^* d} + e^{-\beta^* d}}, \quad \alpha^* = 0.2 \quad \beta^* = 0.6$
Al-Mubaid and Nguyen (2009)	$d_{Mubaid}(c_1, c_2) = \log(1 + len(c_1, c_2) * (depthmax - depth(LCS(c_1, c_2))))$
Pedersen et al. (2007)	$sim_{Path}(c_1, c_2) = \frac{1}{1 + len(c_1, c_2)}$
Sánchez et al. (2012)	$dis_{S\&B}(c_1, c_2) = \log_2 \left(1 + \frac{ \phi(c_1) \setminus \phi(c_2) + \phi(c_2) \setminus \phi(c_1) }{ \phi(c_1) \setminus \phi(c_2) + \phi(c_2) \setminus \phi(c_1) + \phi(c_1) \cap \phi(c_2) } \right)$
	$\phi(a) = \{c \in C \mid a \leq c\}$
	$sim_{Taieb_1}(c_1, c_2) = TermDepth(c_1, c_2) \times TermHypo(c_1, c_2)$
	$TermDepth(c_1, c_2) = \frac{2 \times depth(c_1, c_2)}{depth(c_1) + depth(c_2)}$
	$TermHypo(c_1, c_2) = \frac{2 \times SpecHypo(c_1, c_2)}{SpecHypo(c_1, c_2) + SpecHypo(c_2, c_1)}$
	$SpecHypo(c_1, c_2) = 1 - \frac{\log(HypoValue(c))}{\log(HypoValue(root))}$
Hadj Taieb et al. (2014b)	$HypoValue(c) = \sum_{c' \in HypoInc(c)} P(depth(c'))$
	$P(depth(c')) = \frac{ \{c' \in C \mid depth(c') = depth(c)\} }{ C }$
	$depth(c) = \text{length of the longest ascending path } c \rightarrow \text{root}$
	$HypoInc(c) = \{c' \in C \mid c' \leq c\}$

Table 1: State-of-the-art non IC-based similarity measures evaluated and reproduced in WNSimRep v1.

ment called the root, an Information Content model is a function $IC : C \rightarrow \mathbb{R}^+ \cup \{0\}$, which represents an estimation of the information content for every concept, defined by $IC(c_i) = -\log_2(p(c_i))$, $p(c_i)$ being the occurrence probability of each concept $c_i \in C$. Every IC model must satisfy two further properties: (1) nullity in the root, such that $IC(\Gamma) = 0$, and (2) growing monotonicity from the root to the leaf concepts, such that $\forall c_i \leq c_j \Rightarrow IC(c_i) \geq IC(c_j)$. Once the IC-based measure is chosen, the IC model is mainly responsible for the definition of the notion of similarity and distance between concepts. Other works, such as Pirró and Euzenat (2010), have also proposed intrinsic IC models for semantic relatedness measures which rely on the whole set of semantic relationships encoded into an ontology.

The first known IC model is based on corpus statistics and was introduced by Resnik (1995) and detailed in Resnik (1999). The main drawback of the corpus-based IC models is the difficulty of getting a well-balanced and disambiguated corpus for the estimation of the concept probabilities. To bridge this gap, Seco et al. (2004) introduce the first intrinsic IC model reported in the literature, whose core hypothesis is that the IC models can be directly computed from intrinsic taxonomical features.

1.1 Main motivation

Most ontology-based similarity measures and intrinsic IC models require the computation of different taxonomical features, such as node depths, hyponym sets, node subsumers, Least Common Subsumer (LCS), and subsumed leaves, among others. WordNet is a taxonomy with multiple inheritance, thus, some of these features

are not unambiguously defined, or their computation could be prone to errors. For example, the node depth can be defined as the shortest ascending path length from the node to the root, or the longest ascending path length as defined by Hadj Taieb et al. (2014b). Different definitions of depth also lead us to different values for the LCS concepts. On the other hand, the computation of the hyponym set, subsumed leaves and subsumer set requires a careful counting process to avoid node repetitions, as is already noted in (Seco et al., 2004, §3) when they say “As result of multiple inheritance in some of WordNet’s concepts, caution must be taken so that each distinct hyponym is considered only once”. Another potential source of error is the ambiguity in the definition and notation of some IC models and similarity measures. For example, Zhou et al. (2008b) define the root depth as 1, while the standard convention in graph theory is 0. Most authors define the hyponym set as the descendant node set without including the base node itself. However, in Hadj Taieb et al. (2014b), the hyponym set also includes the base concept. In addition, we find works that do not detail the IC models used in their experiments, or how these IC models were built. Finally, many recent hybrid-type measures also require the computation of the length of the shortest path between concepts. These sources of ambiguity and difficulty demand a lot of attention to the fine details for replicating most IC models and similarity measures in the literature.

The main motivation of this work is to bridge the lack of a gold standard to assist in the exact replication of ontology-based similarity measures and IC models. In a recent work Lastra-Díaz and García-Serrano (2015b), we find some contradictory results and difficulties in repli-

Reference	Classic IC-based similarity measures
Resnik (1995)	$sim_{Resnik}(c_1, c_2) = IC(MICA(c_1, c_2))$
Jiang and Conrath (1997)	$d_{J\&C}(c_1, c_2) = IC(c_1) + IC(c_2) - 2IC(MICA(c_1, c_2))$ $sim_{J\&C}(c_1, c_2) = 1 - \frac{1}{2}d_{J\&C}(c_1, c_2)$
Lin (1998)	$sim_{Lin}(c_1, c_2) = \frac{2IC(MICA(c_1, c_2))}{IC(c_1) + IC(c_2)}$
Reference	IC-based reformulations of the Tversky similarity measure
Pirró and Seco (2008)	$sim_{P\&S}(c_1, c_2) = \begin{cases} \frac{3IC(MICA(c_1, c_2))}{-IC(c_1) - IC(c_2)} & , \text{ if } c_1 \neq c_2 \\ 1 & , \text{ if } c_1 = c_2 \end{cases}$
Reference	Monotone transformations of classic IC-based similarity measures
Pirró and Euzenat (2010)	$sim_{FaITH}(c_1, c_2) = \frac{IC(MICA(c_1, c_2))}{IC(c_1) + IC(c_2) - IC(MICA(c_1, c_2))}$
Meng and Gu (2012)	$sim_{Meng}(c_1, c_2) = e^{sim_{Lin}(c_1, c_2)} - 1 = e^{\frac{2IC(MICA(c_1, c_2))}{IC(c_1) + IC(c_2)}} - 1$
Garla and Brandt (2012)	$sim_{path_IC}(c_1, c_2) = \frac{1}{1 + d_{J\&C}(c_1, c_2)}$
Lastra-Díaz and García-Serrano (2015b)	$sim_{cosJ\&C}(c_1, c_2) = 1 - \cos\left(\frac{\pi}{2} \left(1 - \frac{d_{J\&C}(c_1, c_2)}{2 * max_{d_{J\&C}}}\right)\right)$ $max_{d_{J\&C}} = \max_{c \in Leaves(C)} \{IC(c)\}$
Reference	Hybrid IC-based measures based on the shortest path length
Li et al. (2003)	$sim_{Li_s9}(c_1, c_2) = sim_{Li_s4}(c_1, c_2) * \frac{e^{\lambda * IC} - e^{-\lambda * IC}}{e^{\lambda * IC} + e^{-\lambda * IC}}$, $\lambda^* = 0.4$ $IC = MICA(c_1, c_2)$
Zhou et al. (2008b)	$sim_{Zh}(c_1, c_2) = 1 - k \times \left(\frac{\log(len(c_1, c_2) + 1)}{\log(2 * max_{c \in T}\{depth(c)\} - 1)} \right)$ $-\frac{1}{2}(1 - k) \times d_{J\&C}(c_1, c_2)$ $k^* = \frac{1}{2}$ by default
Meng et al. (2014)	$sim_{Meng2014}(c_1, c_2) = sim_{Lin}(c_1, c_2) \left(\frac{1 - e^{-k * len(c_1, c_2)}}{e^{-k * len(c_1, c_2)}} \right)$, $k^* = 0.08$ $sim_{Gao}(c_1, c_2) = e^{-\alpha L(c_1, c_2)}$, $\alpha^* = 0.15$ and $\beta^* = 2.05$
Gao et al. (2015)	$L(c_1, c_2) = wt(c_1, c_2) * len(c_1, c_2)$ $wt = \begin{cases} \left(\frac{1 + IC(MICA(c_1, c_2))}{IC(MICA(c_1, c_2))} \right)^\beta & , IC(MICA(c_1, c_2)) \geq 1 \\ 2^\beta & , 1 > IC(MICA(c_1, c_2)) \geq 0 \end{cases}$
Lastra-Díaz and García-Serrano (2015b)	$sim_{coswJ\&C}(c_1, c_2) = 1 - \cos\left(\frac{\pi}{2} \left(1 - \frac{d_{wJ\&C}(c_1, c_2)}{2 * max_{d_{J\&C}}}\right)\right)$ $d_{wJ\&C}(c_1, c_2) = \min_{\forall \alpha \in Paths(c_1, c_2)} \left\{ \sum_{e_{ij} \in \alpha} w(e_{ij}) \right\}$ $w(e_{ij}) = \begin{cases} -\log_2(p(c_i c_j)) & , \text{ if } p(c_i c_j) \text{ are known} \\ IC(c_i) - IC(c_j) & , \text{ otherwise} \end{cases}$

Table 2: Definition of the state-of-the-art IC-based similarity measures evaluated and reproduced in WNSimRep v1.

cating previous methods and experiments reported in the literature. These reproducibility problems were confirmed in another subsequent work, such as [Lastra-Díaz and García-Serrano \(2015a\)](#), whilst new contradictory results are reported by [Lastra-Díaz and García-Serrano \(2016b\)](#). Several replication problems were solved with the kind support of most authors. However, we were not able to confirm all previous results, whilst others could not be reproduced through lack of information. As we have already explained, many taxonomical features are ambiguously defined or prone to errors. Thus, all the aforementioned facts lead us to conclude that the exact replication of ontology-based similarity measures and IC models is a hard task, and not exempt from risk. Therefore, it follows that it is urgent and desirable to set of a gold standard for this taxonomical information in order to support the exact replication of the methods reported in the literature.

Most works introducing similarity measures or IC models during the last decade have only implemented or evaluated classic IC-based similarity measures, such as the Resnik, Lin and Jiang-Conrath measures, avoiding the replication of IC models and similarity measures introduced by other researchers. Some works have not included all the details of their methods, or the experimental setup to obtain the published results, thus, preventing their reproducibility. Most works have copied results published by others. This latter fact has prevented the valuable confirmation of previous methods and results reported in the literature, which is an essential feature of science. [Pedersen \(2008a\)](#), and subsequently [Fokkens et al. \(2013\)](#), warn of the need to reproduce and validate previous methods and results reported in the literature, a suggestion that we subscribe to in our aforementioned works, where we also warn of finding some contradictory results. This replication problem is especially significant in the current state of the problem, in which there is no convincing winner within the family of intrinsic IC-based similarity measures and the performance margin is very narrow, as concluded in our aforementioned works and this work. In addition, [Pedersen \(2008a\)](#) also warns of the need of releasing the software developed by the authors of new methods and experiments reported in the literature with the aim of allowing their reproducibility. Following the ideas from Pedersen, the main aim of our main aforementioned companion paper is to introduce and making publicly available our aforementioned software library called HESML, together with a set of reproducible experiments based on ReProZip, [Chirigati et al. \(2013\)](#).

1.2 Research problem and contributions

First aim of this paper is to propose an open framework to assist in the exact replication of most of the intrinsic and corpus-based IC models, intrinsic and IC-based similarity measures, and similarity benchmarks reported in the literature. A second aim is that any further taxonomical feature or taxonomy-based function used by any intrinsic similarity measure or IC model can be represented, at least partially, within the proposed framework. Our final aim for the replication framework de-

scribed herein is to encourage the publication of similar datasets within the research community as a means of improving the reproducibility of ontology-based semantic similarity measures and IC models.

In order to reach the aforementioned aims, we introduce herein a replication framework implemented as a large accompanying dataset of intrinsic and corpus-based Information Content (IC) models in WordNet 3.0. The replication dataset is called *WNSimRep v1* and it is enriched with the most common taxonomical features used in the computation of similarity measures and intrinsic IC models. Despite *WNSimRep v1* is based on WordNet 3.0, the proposed framework could be adapted and extended to any type of base ontology, or intrinsic similarity measure.

The main contribution of this work is the accompanying replication dataset called *WNSimRep v1* which is publicly available at [Lastra-Díaz and García-Serrano \(2016d\)](#). *WNSimRep v1* includes three different types of data files: (1) node-valued IC data files with taxonomical features, (2) edge-valued IC data files with the conditional probability between child and parent concepts, and (3) synset-pair-valued data files with taxonomical features and IC-based similarity measures for the synset pairs derived from the classic RG65 benchmark introduced by [Rubenstein and Goodenough \(1965\)](#). The dataset includes 22 intrinsic IC models, 8 corpus-based IC models based on the Resnik method, 8 corpus-based IC models based on the well-founded *CondProbCorpus* IC model, and 8 corpus-based IC model based on the *CondProbRefCorpus*, which have been evaluated with 22 similarity measures. In addition, the synset-pair-valued data files include the similarity values for all ontology-based similarity measures shown in tables 1 and 2. All the corpus-based IC models are derived from the family of “*add1.dat” WordNet-based frequency files included in the [Pedersen \(2008b\)](#) dataset, which is a dataset of corpus-based files created for a series of papers on similarity measures in WordNet, such as [Patwardhan and Pedersen \(2006\)](#) and [Pedersen \(2010\)](#). The dataset includes all the IC models and similarity measures evaluated in a series of word similarity benchmarks introduced by the authors in [Lastra-Díaz and García-Serrano \(2015b\)](#), [Lastra-Díaz and García-Serrano \(2015a\)](#) and [Lastra-Díaz and García-Serrano \(2016a\)](#).

The rest of the work is structured as follows. Section 2 introduces the ontology-based semantic similarity measures and IC models that have been evaluated and included in the *WNSimRep v1* dataset. Section 3 introduces our replication framework for similarity measures and IC models in WordNet. Section 4 details the licensing information of the *WNSimRep v1* dataset. Finally, we summarize our conclusions and future work.

2 Similarity measures and IC models included

The family of classic ontology-based similarity measures based on IC models is made up by the pioneering work of [Resnik \(1995\)](#), and the subsequent similarity mea-

Reference	Definition of the IC model
Resnik (1999)	$IC_{Resnik} = -\log_2(\hat{p}(c_i))$ $\hat{p}(c_i) = \frac{f(c_i)}{N} = \frac{f(c_i)}{f(\Gamma)}$ $f(c_i) = TF(c_i) + IF(c_i) = TF(c_i) + \sum_{\forall c_j c_i \in LA(c_j)} f(c_j)$
Seco et al. (2004)	$IC_{Seco}(c) = 1 - \frac{\log(Hypo(c) +1)}{\log(max_nodes)}$
Zhou et al. (2008a)	$IC_{Zhou}(c) = k \left(1 - \frac{\log(Hypo(c) +1)}{\log(max_nodes)} \right)$ $+ (1-k) \frac{\log(depth(c))}{\log(depth_{max})}, \quad k^* = \frac{1}{2} \text{ (default)}$
Blanchard et al. (2008)	$IC_g(c_i) = -\log_2 \left(\frac{ SubsumedLeaves(c_i) }{maxLeaves} \right)$ $SubsumedLeaves(c_i) = \{c_j \in C \mid c_j \leq_C c_i \wedge c_j \text{ is leaf}\}$
Sánchez et al. (2011)	$IC_{Sánchez2011}(c_i) = -\log_2 \left(\frac{\frac{ Leaves(c_i) }{ subsumers(c_i) } + 1}{maxLeaves + 1} \right)$ $Leaves(c_i) = \{c_j \in C \mid (c_j \leq_C c_i \wedge c_j \neq c_i) \wedge c_j \text{ is leaf}\}$ $subsumers(c_i) = \{c_j \in C \mid c_i \leq_C c_j\}$
Sánchez and Batet (2012)	$IC_{Sánchez2012}(c) = -\log_2 \left(\frac{commonness(c)}{commonness(root)} \right)$ $\begin{cases} commonness(c) = \frac{1}{ Subsmers(c) } & , c \text{ leaf} \\ commonness(c) = \sum_{\forall l \mid l \text{ is leaf and } l < c} commonness(l) & , c \text{ not leaf} \end{cases}$
Meng et al. (2012)	$IC_{Meng}(c) = \frac{\log(depth(c))}{\log(depth_{max})} \times \left(1 - \frac{\log \left(1 + \sum_{a \in Hypo(c)} \frac{1}{depth(a)} \right)}{\log(Node_{max})} \right)$
Yuan et al. (2013)	$IC_{Yuan}(c) = f_{depth}(c) (1 - f_{leaves}(c)) + f_{hyper}(c)$ $\begin{cases} f_{depth}(c) = \frac{\log(depth(c))}{\log(depth_{max})} \\ f_{leaves}(c) = \frac{\log(Leaves(c) +1)}{\log(Leaves_{max}+1)} \\ f_{hyper}(c) = \frac{\log(Hyper(c) +1)}{\log(Node_{max})} \end{cases}$
Hadj Taieb et al. (2014a)	$IC_{Tajieb}(c) = \left(\sum_{a \in HyperInc(c)} Score(a) \right) \times AvgDepth(c)$ $AvgDepth(c) = \frac{1}{ HyperInc(c) } \times \sum_{c' \in HyperInc(c)} depth(c')$ $Score(c) = \left(\sum_{c' \in DirectHyper(c)} \frac{depth(c')}{ HypoInc(c') } \right) \times HypoInc(c) $ $HypoInc(c) = \{a \in C \mid a \leq c\}$ $HyperInc(c) = \{a \in C \mid c \leq a\}$
Adhikari et al. (2015)	$IC_{Adhikari}(c) = \frac{\log(depth(c)+1)}{\log(depth_{max}+1)} \times \left(1 - \log \left(\frac{\frac{ Leaves(c) \times nmih(c) }{ subsmers'(c) } + 1}{ subsmers'(c) } \right) \right)$ $\times \left(1 - \frac{\log \left(1 + \sum_{a \in Hypo(c)} \frac{1}{depth(a)} \right)}{\log(Node_{max})} \right)$ $subsmers'(c) = subsmers(c) \cup \{c\}$

Table 3: State-of-the-art Information Content models evaluated and reproduced in WNSimRep v1.

sures proposed by Jiang and Conrath (1997) and Lin (1998). In turn, the more recent IC-based similarity measures can be divided into three subgroups as follows: (1) a first group of IC-based similarity measures based on the reformulation strategies between different approaches, such as the IC-based reformulations of the Tversky measure in Pirró (2009) and Pirró and Euzenat (2010), as well as the IC-based reformulation of most edge-counting methods introduced by Sánchez and Batet (2011); (2) a second group of IC-based similarity measures based on a monotone transformation of any classic IC-based similarity measure, such as the exponential-like transformations of the Lin (1998) measure introduced by Meng and Gu (2012) and Pirró and Euzenat (2010), the reciprocal of the Jiang-Conrath distance introduced by Garla and Brandt (2012), and another cosine-based normalization of the Jiang-Conrath distance introduced by Lastra-Díaz and García-Serrano (2015b); and finally, (3) a third group that we call hybrid or path-based IC-based similarity measures, which is defined by those measures that make up an IC model with any function based on the length of the shortest path between concepts, such as the pioneering work of Li et al. (2003), and other subsequent works such as Zhou et al. (2008a), Meng et al. (2014), Gao et al. (2015), and two weighted IC-based similarity measures introduced by Lastra-Díaz and García-Serrano (2015b). Tables 1 and 2 show the definition of the non IC-based similarity measures and IC-based similarity measures that have been evaluated and reproduced in *WNSimRep v1* respectively.

On the other hand, since the pioneering work on intrinsic IC models of (Seco et al., 2004, §3), the development of intrinsic IC models has become one of the mainstreams of research in the area. Among the main intrinsic IC models proposed in the literature, we find the proposals by Zhou et al. (2008a), Sebti and Barfroush (2008), Blanchard et al. (2008), Sánchez et al. (2011), Sánchez and Batet (2012), Yuan et al. (2013), Hadj Taieb et al. (2014a), Lastra-Díaz and García-Serrano (2015a), Adhikari et al. (2015) and Lastra-Díaz and García-Serrano (2016a). Tables 3, 4 and 5 show the definition of the state-of-the-art IC models evaluated and reproduced in the *WNSimRep v1* dataset, which includes the evaluation of all known intrinsic IC models based on WordNet, with the only exception of the IC models introduced by Harispe et al. (2015) and Ben Aouicha et al. (2016).

For a detailed review of the aforementioned methods, we refer the reader to our two previous works on IC-based similarity measures Lastra-Díaz and García-Serrano (2015b) and IC models Lastra-Díaz and García-Serrano (2015a), as well as the recent book by Harispe et al. (2015).

The *WNSimRep v1* dataset is automatically built using the HESML software library introduced in our main paper, Lastra-Díaz and García-Serrano (2016b), which is publicly available at Lastra-Díaz and García-Serrano (2016c). HESML has been developed by the authors in order to replicate all methods reported in the literature and to solve several drawbacks in other publicly available software libraries, such as those introduced by Ped-

ersen et al. (2004) and Harispe et al. (2014). The automatic method used to create the *WNSimRep v1* dataset is described in our aforementioned paper, and it could be also used to generate similar datasets for other similarity measures and IC models on any word similarity benchmark.

IC model	Definition of the IC model
CPHypo	$IC_{CPHypo}(c_i) = -\log_2(p_{Hypo}(c_i))$ $p_{Hypo}(c_i c_j) = \frac{1}{\sum_{\forall c_k c_j \in LA(c_k)} (Hypo(c_k) +1)}$
CPUrif	$IC_{CPUrif}(c_i) = -\log_2(p_{Uniform}(c_i))$ $p_{Uniform}(c_i c_j) = \frac{1}{ children(c_j) }$
CPLeaves	$IC_{CPLeaves}(c_i) = -\log_2(p_{Leaves}(c_i))$ $p_{Leaves}(c_i c_j) = \frac{1}{\sum_{\forall c_k c_j \in LA(c_k)} (Leaves(c_k) +1)}$
CPCorpus	$IC_{CondProbCorpus}(c_i) = -\log_2(p(c_i))$ $p_{corpus}(c_i c_j) = \frac{\max\{1, f(c_i)\}}{\sum_{\forall c_k c_j \in LA(c_k)} \max\{1, f(c_k)\}}$
CPLog	$IC_{CPLog}(c_i) = -\log_2(p_{Log}(c_i))$ $p_{Log}(c_i c_j) = \varphi_l(x) \circ p_{Hypo}(c_i c_j)$ $\varphi_l(x:k) = \frac{1}{1+e^{-k(x-\frac{1}{2})}}, \quad k^* = 8$
CPCosine	$IC_{CPCosine}(c_i) = -\log_2(p_{Cos}(c_i))$ $p_{Cos}(c_i c_j) = \varphi_c(x) \circ p_{Hypo}(c_i c_j)$ $\varphi_c(x) = 1 - \cos\left(\frac{\pi}{2}x\right)$

Table 4: State-of-the-art IC models introduced by Lastra-Díaz and García-Serrano (2015a) which are evaluated and reproduced in *WNSimRep v1*.

3 The replication framework

In this section, we introduce a framework which defines the data that is needed to reproduce the most significant ontology-based similarity measures, as well as all the known WordNet-based intrinsic and corpus-based IC models with the only exception of the IC models introduced by Harispe et al. (2015) and Ben Aouicha et al. (2016). In order to replicate these mathematical models it is convenient to have a data collection that allows the validation of any novel code implementation of the models and measures, and the reproduction of the experiments carried-out by the authors. Any researcher or practitioner could use the data in the *WNSimRep v1* dataset for the verification of the computation of any taxonomical feature in their own code implementation. The aim of the *WNSimRep v1* dataset is to become a gold standard for most taxonomical features involved into the definition of most IC models and ontology-based similarity measures reported in the literature. In addition, *WNSimRep v1* provides the full data of most IC models, as well as the similarity values reported by most ontology-based similarity measures in the RG65 dataset. Our final aim is that the authors of new methods reported in the literature publish similar datasets in order to encourage the exact replication of their methods and experiments.

To achieve the aims of this work, we define three different types of data file for each base taxonomy and

IC model. These files encode three types of function: node-valued functions, edge-valued functions and binary synset-valued functions. In this way, the framework for each base taxonomy and IC model is defined as follows: (1) a *node-valued data file* which contains an explicit representation of the base taxonomy together with the node-based IC values, plus a collection of node-valued taxonomical features that should include all the features required to replicate the target IC model or similarity measure; (2) an *edge-valued data file* for the edge-based IC models, such as our *CondProb** and *CondProbRef** families, which contains the conditional probabilities for each taxonomy edge, the edge-based IC values and the edge-based weight whenever necessary; and (3) a *synset-valued data file* for each intrinsic IC model in tables 3, 4 and 5, which includes a set of IC-based similarity measures, and the most common taxonomical features from among the synsets associated to the word pairs included in the RG65 benchmark, such as the shortest path length, lowest common subsumer (LCS) and most informative common ancestor (MICA). This latest data file could be extended to include all pairwise distances between synsets in WordNet, however its computation time would be excessive. In addition, the *WNSimRep v1* also includes a collection of *synset-valued data files* for every non IC-based similarity measure in table 1. The aim is to provide all the information necessary in the validation process of any novel code implementation of these mathematical models, or the reproduction of experiments reported in the literature. For the sake of completeness, the *WNSimRep v1* dataset includes the full collection of corpus-based IC models based on the Resnik, *CondProbCorpus* and *CondProbRefCorpus* IC models with the full set of “*add1.dat” WordNet-based frequency files in the Pedersen (2008b) dataset, which were evaluated by Lastra-Díaz and García-Serrano (2015b). Despite the initial framework being proposed for WordNet 3.0 and some known IC models and similarity measures, it can be easily adapted to other similarity measures and base ontologies.

Node-valued data files. Tables 6 and 7 show a list of the intrinsic and corpus-based IC models contained in *WNSimRep v1*. Every file in the dataset contains a table in standard *.csv file format separated by semicolon, which can be directly imported into MS Excel. For each concept $c_i \in C$ within the noun database of WordNet 3.0, there is a row in the node-valued files containing the following information: (1) *synset ID*, (2) *synset words*, (3) *synset ID* of the *parent* nodes, (4) *concept IC value*, (5) *concept probability* whenever it is computed by the IC model, (6) *depth* defined as the shortest ascending path length between the c_i concept and the root concept Γ , (7) *longest depth* defined as the longest ascending path length from c_i to the root, (8) *number of direct child concepts* (direct hyponyms), (9) *number of parent concepts* (direct subsumers), (10) *number of subsumer concepts* excluding the base c_i concept, as defined by $|\{c_j \in C \mid c_i <_C c_j\}|$, (11) *number of hyponym concepts* excluding the base c_i concept, as defined by $|\{c_j \in C \mid c_j <_C c_i\}|$, and (12) *number of leaf subsumed concepts* by the concept c_i , without including it. This

collection of node-valued features could be extended as necessary. For instance, the *HypoValue*(c_i) function defined by Hadj Taieb et al (2014) could be included in the files to assist in its replication.

Edge-valued data files. Table 8 shows the collection of edge-valued data files included in the accompanying dataset. All the IC models within our *CondProb** and *CondProbRef** families are based on the computation of the edge-based conditional probabilities $p(c_i|c_j)$. Thus, for each edge in the WordNet taxonomy there is a row in the data files containing the following set of attributes: (1) *child synset ID*, (2) *parent synset ID*, (3) *conditional probability*, and (4) *edge-based IC weight* as defined by $IC(e_{ij}) = -\log_2(p(c_i|c_j))$.

Synset-valued data files. Table 10 shows the collection of synset-valued data files included in the accompanying dataset. There is one file for each intrinsic IC model shown in table 3. Each row in the data files defines a collection of taxonomical features for each synset pair associated with any word pair in the RG65 dataset. The synset pairs correspond to the Cartesian product between the synsets for each word, thus, the rows are divided into blocks per word pair. Each row includes the following attributes: (1) *synset ID1*, (2) *synset ID2* (3) *length of the shortest path* between the synset pair, (4) *lowest common subsumer (LCS)* between the synset pair based on the minimum depth defined as the shortest ascending path from any node to the root, (5) *lowest common subsumer (longest depth)* between the synset pair based on the maximum depth defined as the longest ascending path from any node to the root, (6) *most informative common ancestor (MICA)*, (7) *MICA IC value*, and finally (8) the *similarity value* for each synset that is returned by each similarity measure shown in table 2.

3.1 How can the *WNSimRep* dataset be used?

The core idea of the framework is to provide enough intermediate data in order to assist the replication process of any IC model or ontology-based similarity measure. First, all the IC models in tables 3, 4 and 5 are explicitly represented in the node-valued data files, thus, any user of the dataset can use these IC models without to implement them by loading into memory the IC values provided by the data files. Second, the *WNSimRep v1* dataset can be used as a gold standard in the validation process of any independent implementation of an intrinsic or corpus-based IC model. For instance, the users can compare the IC values derived from their code implementation with the IC values in the data files, as well as all intermediate taxonomical features involved in their computation. Third, most of the IC models and ontology-based similarity measures in tables 1, 2, 3, 4 and 5, can be directly computed using the node-valued features included in the *WNSimRep v1* dataset. For instance, the Seco et al. (2004), Zhou et al. (2008a) and *CondProbHypo* IC models use the $|Hypo(c)|$ function which is defined by the field (11) within the node-valued data files, thus, any researcher trying to replicate these IC models can compare the *Hypo*(c) values re-

IC model	Definition of the IC model
CondProbRefHyponyms	$IC_{CPRefHypo}(c_i) = -\log_2(p_{Hypo}^*(c_i))$ $p_{Hypo}(c_i c_j) = \frac{1}{\sum_{\forall c_k c_j \in LA(c_k)} (Hypo(c_i) +1)}$
CondProbRefUniform	$IC_{CPRefUni}(c_i) = -\log_2(p_{Uniform}^*(c_i))$ $p_{Uniform}(c_i c_j) = \frac{1}{ children(c_j) }$
CondProbRefLeaves	$IC_{CPRefLea}(c_i) = -\log_2(p_{Leaves}^*(c_i))$ $p_{Leaves}(c_i c_j) = \frac{1}{\sum_{\forall c_k c_j \in LA(c_k)} (Leaves(c_i) +1)}$
CondProbRefLogistic	$IC_{CPRefLog}(c_i) = -\log_2(p_{Log}^*(c_i))$ $p_{Log}(c_i c_j) = \varphi_l(x) \circ p_{Hypo}(c_i c_j)$ $\varphi_l(x : k) = \frac{1}{1+e^{-k(x-\frac{1}{2})}}, \quad k^* = 8$
CondProbRefCosine	$IC_{CPRefCos}(c_i) = -\log_2(p_{Cos}^*(c_i))$ $p_{Cos}(c_i c_j) = \varphi_c(x) \circ p_{Hypo}(c_i c_j)$ $\varphi_c(x) = 1 - \cos\left(\frac{\pi}{2}x\right)$
CondProbRefCorpus	$IC_{CPRefCorpus}(c_i) = -\log_2(p^*(c_i))$ $p_{corpus}(c_i c_j) = \frac{\max\{1, f(c_i)\}}{\sum_{\forall c_k c_j \in LA(c_k)} \max\{1, f(c_k)\}}$
CondProbRefLogisticLeaves	$IC_{CPRefLogLeaves}(c_i) = -\log_2(p_{LogLeaves}^*(c_i))$ $p_{LogLeaves}(c_i c_j) = \varphi_l(x) \circ p_{Leaves}(c_i c_j)$ $\varphi_l(x : k) = \frac{1}{1+e^{-k(x-\frac{1}{2})}}, \quad k^* = 8$
CondProbRefCosineLeaves	$IC_{CPRefCosLeaves}(c_i) = -\log_2(p_{CosLeaves}^*(c_i))$ $p_{CosLeaves}(c_i c_j) = \varphi_c(x) \circ p_{Leaves}(c_i c_j)$ $\varphi_c(x) = 1 - \cos\left(\frac{\pi}{2}x\right)$
CondProbRefLeavesSubsumersRatio	$IC_{CPRefLeaSubRat}(c_i) = -\log_2(p_{LeaSubRat}^*(c_i))$ $p_{LeaSubRat}(c_i c_j) = \frac{\frac{\sigma(c_i)}{\sigma(c_j)}}{\sum_{\forall c_k c_j \in LA(c_k)} \frac{\sigma(c_i)}{\sigma(c_j)}}$ $\sigma(c) = \frac{ Leaves(c) }{ subsumers(c) } + 1$

Table 5: State-of-the-art IC models introduced by [Lastra-Díaz and García-Serrano \(2016a\)](#) which are evaluated and reproduced in WNSimRep v1..

ported by his code implementation with the values provided within the dataset files. Other possibility is to compute the targeted IC model by defining the concept IC values through a formula that uses any available taxonomical feature within the files. The max_nodes value is obtained as $|Hypo(\Gamma)|$ plus 1, using the field (11) for the root concept. The $depth$ function used in the Zhou et al. IC model and measures correspond to the depth function (field 6) plus 1, because the authors define the depth of the root node as 1. The Sánchez et al. (2011) IC model uses the count of subsumed leaves (field 12) and subsumer concepts (field 10), which can be obtained from the node-valued data files. The Yuan et al. IC model uses three different taxonomical features: depth (field 6), subsumed leaves (field 12) and hypernym set counting (subsumer set, field 10). On the other hand, the [Sánchez and Batet \(2012\)](#) IC model requires the computation of the *commonness* function, which is not yet included in our dataset.

The edge-valued data files allow the exact replication of all the well-founded IC models in tables 4 and 5. For example, the conditional probabilities can be computed using the node-valued $|Hypo(c)|$, $|Leaves(c)|$ and $|children(c)|$ functions provided in the accompanying node-valued data files, whose values can then be com-

pared with the conditional probabilities included in the edge-valued data files.

Finally, the synset-based data files are especially helpful in the replication of most similarity measures and benchmarks, because they include the concept-to-concept similarity values returned by the similarity measures in the RG65 dataset, thus, any researcher or practitioner can compare the values returned by his code implementation with the values provided by the dataset. All the IC-based similarity measures in table 2 are already included in the synset-based data files, in which you can obtain the similarity values for each synset pair associated to any word pair in the RG65 dataset. The synset-pair-valued data files include the shortest path length, least common subsumer (LCS) using two depth definitions, and the MICA node and MICA value for each synset pair in the RG65 dataset. This information allows the direct computation of all the similarity measures in tables 1 and 2, with the exception of the Hadj Taieb et al. measure, which requires the *HypoValue(c)* function shown in table 1. However, this data could be provided in a further data file. *WNSimRep v1* is a first try at assisting the research community in the replication of similarity measures and experiments, but it is still open, and it could be updated in the future.

4 Licensing information

The *WNSimRep v1* accompanying dataset is distributed under a Creative Commons By-NC 3.0 license described at creativecommons.org. It means that any user has the right to use or to distribute freely the *WNSimRep v1* for any non commercial use. The users of the accompanying dataset must recognize the authorship of the dataset by citing the main research paper associated to the present work, which is introduced by [Lastra-Díaz and García-Serrano \(2016b\)](#).

Likewise, the *WNSimRep v1* distributes an explicit representation of the WordNet 3.0 taxonomy in their node-valued data files, thus, the users of the dataset must fulfill the licensing requirements of WordNet as are described at wordnet.princeton.edu/wordnet/license/, and they must also cite the WordNet papers introduced by [Miller \(1995\)](#) and [Fellbaum \(1998\)](#), as described at wordnet.princeton.edu/wordnet/citing-wordnet/.

Finally, *WNSimRep v1* also includes a series of corpus-based IC models derived from a subset of the dataset of Wordnet-based frequency files created by [Pedersen \(2008b\)](#), as well as the original source files. Thus, the users of the *WNSimRep v1* dataset must recognize these contributions by citing the papers involved in the development of the aforementioned dataset introduced by [Patwardhan and Pedersen \(2006\)](#) and [Pedersen \(2010\)](#).

5 Conclusions and future work

We have introduced an open framework and dataset to assist in the exact replication of ontology-based similarity measures, IC models and similarity benchmarks reported in the literature. The dataset is publicly available at [Lastra-Díaz and García-Serrano \(2016d\)](#) under a Creative Commons By-NC 3.0 license, and it contains all the IC models and ontology-based similarity measures evaluated in our three previous aforementioned works. As forthcoming activities, we plan to extend the dataset to include more specific features considered in some similarity measures and IC models. In addition, we are carrying-out a comparison study between IC models using the *WNSimRep v1* dataset.

6 Acknowledgements

Ted Pedersen kindly answered all our questions and provided us the WordNet-based frequency files used in our experiments and the corpus-based IC models included in the *WNSimRep* dataset. Mark Hallett checked the proper use of the English. We express our most sincere gratitude to all them. This work has been partially supported by the Spanish VOXPOPULI Project (TIN2013-47090-C3-1-P).

References

Adhikari, A., Singh, S., Dutta, A., Dutta, B., 2015. A novel information theoretic approach for finding semantic similarity in WordNet. In: Proc. of IEEE

International Technical Conference (TENCON-2015). IEEE, Macau, China, pp. 1–4.

Al-Mubaid, H., Nguyen, H. A., Jul. 2009. Measuring Semantic Similarity Between Biomedical Concepts Within Multiple Ontologies. IEEE transactions on systems, man and cybernetics. Part C, Applications and reviews: a publication of the IEEE Systems, Man, and Cybernetics Society 39 (4), 389–398.

Batet, M., Sánchez, D., Valls, A., Feb. 2011. An ontology-based measure to compute semantic similarity in biomedicine. Journal of Biomedical Informatics 44 (1), 118–125.

Ben Aouicha, M., Taieb, M. A. H., Ben Hamadou, A., 28 Mar. 2016. Taxonomy-based information content and wordnet-wiktionary-wikipedia glosses for semantic relatedness. Applied Intelligence, 1–37.

Blanchard, E., Harzallah, M., Kuntz, P., 2008. A generic framework for comparing semantic similarities on a subsumption hierarchy. In: Ghallab, M., Spyropoulos, C. D., Fakotakis, N., Avouris, N. (Eds.), Proceedings of the ECAI. Vol. 178 of Frontiers in Artificial Intelligence and Applications. IOS Press, pp. 20–24.

Chirigati, F., Shasha, D., Freire, J., 2013. Rezip: Using provenance to support computational reproducibility. In: Presented as part of the 5th USENIX Workshop on the Theory and Practice of Provenance. usenix.org.

Fellbaum, C. (Ed.), 1998. WordNet: An Electronic Lexical Database. MIT Press, Cambridge, MA.

Fokkens, A., Van Erp, M., Postma, M., Pedersen, T., Vossen, P., Freire, N., 4 Aug. 2013. Offspring from Reproduction Problems: What Replication Failure Teaches Us. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. ACL, Sofia, Bulgaria, pp. 1691–1701.

Gao, J. B., Zhang, B. W., Chen, X. H., Mar. 2015. A WordNet-based semantic similarity measurement combining edge-counting and information content theory. Engineering Applications of Artificial Intelligence 39 (0), 80–88.

Garla, V. N., Brandt, C., 10 Oct. 2012. Semantic similarity in the biomedical domain: an evaluation across knowledge sources. BMC bioinformatics 13:261.

Hadj Taieb, M. A., Ben Aouicha, M., Ben Hamadou, A., 1 Nov. 2014a. A new semantic relatedness measurement using WordNet features. Knowledge and Information Systems 41 (2), 467–497.

Hadj Taieb, M. A., Ben Aouicha, M., Ben Hamadou, A., Nov. 2014b. Ontology-based approach for measuring semantic similarity. Engineering Applications of Artificial Intelligence 36, 238–261.

Harispe, S., Imoussaten, A., Troussset, F., Montmain, J., 2 Aug. 2015. On the consideration of a bring-to-mind model for computing the Information Content

- of concepts defined into ontologies. In: Proceedings of the IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2015). IEEE, Istanbul, Turkey, pp. 1–8.
- Harispe, S., Ranwez, S., Janaqi, S., Montmain, J., 18 Jun. 2014. The Semantic Measures Library: Assessing Semantic Similarity from Knowledge Representation Analysis. In: Métails, E., Roche, M., Teisseire, M. (Eds.), Proc. of the 19th International Conference on Applications of Natural Language to Information Systems (NLDB 2014). Vol. 8455 of LNCS. Springer, Montpellier, France, pp. 254–257.
- Jiang, J. J., Conrath, D. W., 1997. Semantic similarity based on corpus statistics and lexical taxonomy. In: Proceedings of International Conference Research on Computational Linguistics (ROCLING X). pp. 19–33.
- Lastra-Díaz, J. J., García-Serrano, A., Nov. 2015a. A new family of information content models with an experimental survey on WordNet. Knowledge-Based Systems 89, 509–526.
- Lastra-Díaz, J. J., García-Serrano, A., Nov. 2015b. A novel family of IC-based similarity measures with a detailed experimental survey on WordNet. Engineering Applications of Artificial Intelligence Journal 46, 140–153.
- Lastra-Díaz, J. J., García-Serrano, A., Jul. 2016a. A refinement of the well-founded Information Content models with a very detailed experimental survey on WordNet. Tech. Rep. TR-2016-01, Department of Computer Languages and Systems. Universidad Nacional de Educación a Distancia (UNED), <http://e-spacio.uned.es/fez/view/bibliuned:DptoLSI-ETSI-Informes-Jlastra-refinement>.
- Lastra-Díaz, J. J., García-Serrano, A., 2016b. HESML: a scalable ontology-based semantic similarity measures library with a set of reproducible experiments and a replication dataset. To appear in Information Systems.
- Lastra-Díaz, J. J., García-Serrano, A., 2016c. HESML VIR1 Java software library of ontology-based semantic similarity measures and information content models. Mendeley Data v1, <http://dx.doi.org/10.17632/t87s78dg78.1>.
- Lastra-Díaz, J. J., García-Serrano, A., 2016d. WN-SimRep: a framework and replication dataset for ontology-based semantic similarity measures and information content models. Mendeley Data v1, <http://dx.doi.org/10.17632/mpr2m8pycs.1>.
- Leacock, C., Chodorow, M., 1998. Combining local context and WordNet similarity for word sense identification. In: Fellbaum, C. (Ed.), WordNet: An electronic lexical database. MIT Press, Ch. 11, pp. 265–283.
- Li, Y., Bandar, Z. A., McLean, D., 2003. An approach for measuring semantic similarity between words using multiple information sources. IEEE Transactions on Knowledge and Data Engineering 15 (4), 871–882.
- Lin, D., 1998. An information-theoretic definition of similarity. In: Proceedings of the 15th International Conference on Machine Learning. Vol. 98. Madison, WI, pp. 296–304.
- Meng, L., Gu, J., 2012. A New Model for Measuring Word Sense Similarity in WordNet. In: Proceedings of the 4th International Conference on Advanced Communication and Networking, ASTL. Vol. 14. pp. 18–23.
- Meng, L., Gu, J., Zhou, Z., Sep. 2012. A new model of information content based on concept’s topology for measuring semantic similarity in WordNet. International Journal of Grid and Distributed Computing 5 (3), 81–93.
- Meng, L., Huang, R., Gu, J., Jun. 2014. Measuring Semantic Similarity of Word Pairs Using Path and Information Content. International Journal of Future Generation Communication & Networking 7 (3), 183–194.
- Miller, G. A., 1995. WordNet: A Lexical Database for English. Communications of the ACM 38 (11), 39–41.
- Patwardhan, S., Pedersen, T., 2006. Using WordNet-based context vectors to estimate the semantic relatedness of concepts. In: Proceedings of the EACL 2006 Workshop Making Sense of Sense-Bringing Computational Linguistics and Psycholinguistics Together. Vol. 1501. Trento, Italy, pp. 1–8.
- Pedersen, T., 2008a. Empiricism Is Not a Matter of Faith. Computational Linguistics 34 (3), 465–470.
- Pedersen, T., 2008b. WordNet-InfoContent-3.0.tar dataset repository. https://www.researchgate.net/publication/273885902_WordNet-InfoContent-3.0.tar.
- Pedersen, T., 2010. Information Content Measures of Semantic Similarity Perform Better Without Sense-tagged Text. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. HLT ’10. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 329–332.
- Pedersen, T., Pakhomov, S. V. S., Patwardhan, S., Chute, C. G., Jun. 2007. Measures of semantic similarity and relatedness in the biomedical domain. Journal of biomedical informatics 40 (3), 288–299.
- Pedersen, T., Patwardhan, S., Michelizzi, J., 2004. WordNet::Similarity: Measuring the Relatedness of Concepts. In: Demonstration Papers at HLT-NAACL 2004. HLT-NAACL–Demonstrations ’04. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 38–41.
- Pirró, G., Nov. 2009. A semantic similarity metric combining features and intrinsic information content. Data & Knowledge Engineering 68 (11), 1289–1308.

- Pirró, G., Euzenat, J., 7 Nov. 2010. A Feature and Information Theoretic Framework for Semantic Similarity and Relatedness. In: Patel-Schneider, P. F., Pan, Y., Hitzler, P., Mika, P., Zhang, L., Pan, J. Z., Horrocks, I., Glimm, B. (Eds.), Proc. of the 9th International Semantic Web Conference, ISWC 2010. Vol. 6496 of LNCS. Springer, Shangai, China, pp. 615–630.
- Pirró, G., Seco, N., Jan. 2008. Design, Implementation and Evaluation of a New Semantic Similarity Metric Combining Features and Intrinsic Information Content. In: Meersman, R., Tari, Z. (Eds.), On the Move to Meaningful Internet Systems: OTM 2008. Vol. 5332 of LNCS. Springer, pp. 1271–1288.
- Rada, R., Mili, H., Bicknell, E., Blettner, M., 1989. Development and application of a metric on semantic nets. IEEE Transactions on Systems, Man, and Cybernetics 19 (1), 17–30.
- Resnik, P., 20 Aug. 1995. Using information content to evaluate semantic similarity in a taxonomy. In: Proc. of the International Joint Conferences on Artificial Intelligence (IJCAI 1995). Vol. 1. Montreal, Canada, pp. 448–453.
- Resnik, P., Jul. 1999. Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. Journal of Artificial Intelligence Research 11, 95–130.
- Rubenstein, H., Goodenough, J. B., Oct. 1965. Contextual Correlates of Synonymy. Communications of the ACM 8 (10), 627–633.
- Sánchez, D., Batet, M., Oct. 2011. Semantic similarity estimation in the biomedical domain: an ontology-based information-theoretic perspective. Journal of biomedical informatics 44 (5), 749–759.
- Sánchez, D., Batet, M., 2012. A new model to compute the information content of concepts from taxonomic knowledge. International Journal on Semantic Web and Information Systems (ISWIS) 8 (2), 34–50.
- Sánchez, D., Batet, M., Isern, D., Mar. 2011. Ontology-based information content computation. Knowledge-Based Systems 24 (2), 297–303.
- Sánchez, D., Batet, M., Isern, D., Valls, A., 2012. Ontology-based semantic similarity: A new feature-based approach. Expert Systems with Applications 39, 7718–7728.
- Sebti, A., Barfroush, A. A., Oct. 2008. A new word sense similarity measure in WordNet. In: Proc. of the International Multiconference on Computer Science and Information Technology, IMCSIT 2008. IEEE, pp. 369–373.
- Seco, N., Veale, T., Hayes, J., 2004. An intrinsic information content metric for semantic similarity in WordNet. In: López de Mántaras, R., Saitta, L. (Eds.), Proceedings of the 16th European Conference on Artificial Intelligence (ECAI). Vol. 16. IOS Press, Valencia, Spain, pp. 1089–1094.
- Tversky, A., Jul. 1977. Features of similarity. Psychological Review 84 (4), 327–352.
- Wu, Z., Palmer, M., 1994. Verbs Semantics and Lexical Selection. In: Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics. ACL ’94. ACL, Stroudsburg, PA, USA, pp. 133–138.
- Yuan, Q., Yu, Z., Wang, K., Dec. 2013. A New Model of Information Content for Measuring the Semantic Similarity between Concepts. In: Proc. of the International Conference on Cloud Computing and Big Data (CloudCom-Asia 2013). IEEE Computer Society, pp. 141–146.
- Zhou, Z., Wang, Y., Gu, J., 2008a. A new model of information content for semantic similarity in WordNet. In: Proc. of the Second International Conference on Future Generation Communication and Networking Symposia (FGCNS’08). Vol. 3. IEEE, pp. 85–89.
- Zhou, Z., Wang, Y., Gu, J., Nov. 2008b. New model of semantic similarity measuring in WordNet. In: Proc. of the 3rd International Conference on Intelligent System and Knowledge Engineering (ISKE 2008). Vol. 1. IEEE, pp. 256–261.

7 Appendix: WNSimRep v1 dataset files.

Tables 6 to 11 show the current files included in the *WN-SimRep v1* dataset

IC model reference	Node-valued intrinsic IC model files
Seco et al. (2004)	WNSimRep_ICmodel_Seco
Blanchard et al. (2008)	WNSimRep_ICmodel_Blanchard
Zhou et al. (2008a)	WNSimRep_ICmodel_Zhou
Sánchez et al. (2011)	WNSimRep_ICmodel_Sanchez2011
Sánchez and Batet (2012)	WNSimRep_ICmodel_Sanchez2012
Meng et al. (2012)	WNSimRep_ICmodel_Meng
Yuan et al. (2013)	WNSimRep_ICmodel_Yuan
Hadj Taieb et al. (2014a)	WNSimRep_ICmodel_Hadj_Taieb
Adhikari et al. (2015)	WNSimRep_ICmodel_Adhikari
	WNSimRep_ICmodel_CondProbHyponyms
	WNSimRep_ICmodel_CondProbUniform
Lastra-Díaz and García-Serrano (2015a)	WNSimRep_ICmodel_CondProbLeaves
	WNSimRep_ICmodel_CondProbLogistic
	WNSimRep_ICmodel_CondProbCosine
	WNSimRep_ICmodel_CondProbRefHyponyms
	WNSimRep_ICmodel_CondProbRefUniform
	WNSimRep_ICmodel_CondProbRefLeaves
	WNSimRep_ICmodel_CondProbRefLogistic
Lastra-Díaz and García-Serrano (2016a)	WNSimRep_ICmodel_CondProbRefCosine
	WNSimRep_ICmodel_CondProbRefLogisticLeaves
	WNSimRep_ICmodel_CondProbRefCosineLeaves
	WNSimRep_ICmodel_CondProbRefLeavesSubsumersRatio

Table 6: Intrinsic IC model files included in the WNSimRep v1 dataset.

Node-valued files derived from the CondProbCorpus IC model, Lastra-Díaz and García-Serrano (2015a)	
CondProbCorpus	WNSimRep_ICmodel_CondProbCorpus_ic-bnc-resnik-add1
	WNSimRep_ICmodel_CondProbCorpus_ic-brown-resnik-add1
	WNSimRep_ICmodel_CondProbCorpus_ic-semcor-add1
	WNSimRep_ICmodel_CondProbCorpus_ic-semcorraw-add1
	WNSimRep_ICmodel_CondProbCorpus_ic-semcorraw-resnik-add1
	WNSimRep_ICmodel_CondProbCorpus_ic-shaks-resnink-add1
	WNSimRep_ICmodel_CondProbCorpus_ic-treebank-add1
	WNSimRep_ICmodel_CondProbCorpus_ic-treebank-resnik-add1
Node-valued files derived from the Resnik IC model, Resnik (1999)	
Resnik	WNSimRep_ICmodel_ResnikMethod_ic-bnc-resnik-add1
	WNSimRep_ICmodel_ResnikMethod_ic-brown-resnik-add1
	WNSimRep_ICmodel_ResnikMethod_ic-semcor-add1
	WNSimRep_ICmodel_ResnikMethod_ic-semcorraw-add1
	WNSimRep_ICmodel_ResnikMethod_ic-semcorraw-resnik-add1
	WNSimRep_ICmodel_ResnikMethod_ic-shaks-resnink-add1
	WNSimRep_ICmodel_ResnikMethod_ic-treebank-add1
	WNSimRep_ICmodel_ResnikMethod_ic-treebank-resnik-add1
Node-valued files derived from the CondRefProbCorpus IC model, Lastra-Díaz and García-Serrano (2016a)	
CondProbRefCorpus	WNSimRep_ICmodel_CondProbRefCorpus_ic-bnc-resnik-add1
	WNSimRep_ICmodel_CondProbRefCorpus_ic-brown-resnik-add1
	WNSimRep_ICmodel_CondProbRefCorpus_ic-semcor-add1
	WNSimRep_ICmodel_CondProbRefCorpus_ic-semcorraw-add1
	WNSimRep_ICmodel_CondProbRefCorpus_ic-semcorraw-resnik-add1
	WNSimRep_ICmodel_CondProbRefCorpus_ic-shaks-resnink-add1
	WNSimRep_ICmodel_CondProbvCorpus_ic-treebank-add1
	WNSimRep_ICmodel_CondProbRefCorpus_ic-treebank-resnik-add1

Table 7: Corpus-based IC model files included in the WNSimRep v1 dataset. All the IC models have been computed by using the WordNet-based frequency files included in [Pedersen \(2008b\)](#).

CondProbCorpus IC model files introduced by [Lastra-Díaz and García-Serrano \(2015a\)](#)

WNSimRep_EdgeInfo_CondProbCorpus_ic-bnc-resnik-add1
WNSimRep_EdgeInfo_CondProbCorpus_ic-brown-resnik-add1
WNSimRep_EdgeInfo_CondProbCorpus_ic-semcor-add1
WNSimRep_EdgeInfo_CondProbCorpus_ic-semcorraw-add1
WNSimRep_EdgeInfo_CondProbCorpus_ic-semcorraw-resnik-add1
WNSimRep_EdgeInfo_CondProbCorpus_ic-shaks-resnink-add1
WNSimRep_EdgeInfo_CondProbCorpus_ic-treebank-add1
WNSimRep_EdgeInfo_CondProbCorpus_ic-treebank-resnik-add1

CondProbRefCorpus IC model files introduced by [Lastra-Díaz and García-Serrano \(2016a\)](#)

WNSimRep_EdgeInfo_CondProbRefCorpus_ic-bnc-resnik-add1
WNSimRep_EdgeInfo_CondProbRefCorpus_ic-brown-resnik-add1
WNSimRep_EdgeInfo_CondProbRefCorpus_ic-semcor-add1
WNSimRep_EdgeInfo_CondProbRefCorpus_ic-semcorraw-add1
WNSimRep_EdgeInfo_CondProbRefCorpus_ic-semcorraw-resnik-add1
WNSimRep_EdgeInfo_CondProbRefCorpus_ic-shaks-resnink-add1
WNSimRep_EdgeInfo_CondProbRefCorpus_ic-treebank-add1
WNSimRep_EdgeInfo_CondProbRefCorpus_ic-treebank-resnik-add1

Table 8: Edge-valued corpus-based IC model files included in the accompanying dataset. Each file contains several edge-valued features for each edge of the taxonomy.

Intrinsic IC model files.introduced by [Lastra-Díaz and García-Serrano \(2015a\)](#)

CondProbHyponyms	WNSimRep_EdgeInfo_CondProbHyponyms
CondProbUniform	WNSimRep_EdgeInfo_CondProbUniform
CondProbLeaves	WNSimRep_EdgeInfo_CondProbLeaves
CondProbLogistic	WNSimRep_EdgeInfo_CondProbLogistic
CondProbCosine	WNSimRep_EdgeInfo_CondProbCosine

Intrinsic IC model files.introduced by [Lastra-Díaz and García-Serrano \(2016a\)](#)

CondProbRefHyponyms	WNSimRep_EdgeInfo_CondProbRefHyponyms
CondProbRefUniform	WNSimRep_EdgeInfo_CondProbRefUniform
CondProbRefLeaves	WNSimRep_EdgeInfo_CondProbRefLeaves
CondProbRefLogistic	WNSimRep_EdgeInfo_CondProbRefLogistic
CondProbRefCosine	WNSimRep_EdgeInfo_CondProbRefCosine
CPreLogisticLeaves	WNSimRep_EdgeInfo_CondProbRefLogisticLeaves
CPreCosineLeaves	WNSimRep_EdgeInfo_CondProbRefCosineLeaves
CPreLeaSubRatio	WNSimRep_EdgeInfo_CondProbRefLeavesSubsumersRatio

Table 9: Edge-valued intrinsic IC model files included in the accompanying dataset. Each file contains several edge-valued features for each edge of the taxonomy.

IC model	Synset pair-valued IC model files (RG65)
Seco et al. (2004)	WNSimRep_SynsetPairs_Seco
Blanchard et al. (2008)	WNSimRep_SynsetPairs_Blanchard
Zhou et al. (2008a)	WNSimRep_SynsetPairs_Zhou
Sánchez et al. (2011)	WNSimRep_SynsetPairs_Sanchez2011
Sánchez and Batet (2012)	WNSimRep_SynsetPairs_Sanchez2012
Meng et al. (2012)	WNSimRep_SynsetPairs_Meng
Yuan et al. (2013)	WNSimRep_SynsetPairs_Yuan
Adhikari et al. (2015)	WNSimRep_SynsetPairs_Adhikari
Hadj Taieb et al. (2014a)	WNSimRep_SynsetPairs_HadjTaieb
Lastra-Díaz and García-Serrano (2015a)	WNSimRep_SynsetPairs_CondProbHyponyms
	WNSimRep_SynsetPairs_CondProbUniform
	WNSimRep_SynsetPairs_CondProbLeaves
	WNSimRep_SynsetPairs_CondProbLogistic
	WNSimRep_SynsetPairs_CondProbCosine
Lastra-Díaz and García-Serrano (2016a)	WNSimRep_SynsetPairs_CondProbRefHyponyms
	WNSimRep_SynsetPairs_CondProbRefUniform
	WNSimRep_SynsetPairs_CondProbRefLeaves
	WNSimRep_SynsetPairs_CondProbRefLogistic
	WNSimRep_SynsetPairs_CondProbRefCosine
	WNSimRep_SynsetPairs_CondProbRefLogistic
	WNSimRep_SynsetPairs_CondProbRefCosine
	WNSimRep_SynsetPairs_CondProbRefLeavesSubsumersRatio

Table 10: Synset-valued data files based on the RG65 dataset and the intrinsic IC model files included in the WNSimRep v1 dataset. Each file contains a set of taxonomical features and the degree of similarity between the two synsets corresponding to each synset pair between two input words. The files contain the similarity values reported by all IC-based similarity measures in table 2.

Similarity measures	Synset pair-valued file (RG65)
Rada et al. (1989)	WNSimRep_SynsetPairs_nonIC_based_measures
Wu and Palmer (1994)	
Leacock and Chodorow (1998)	
Li et al. (2003)	
Li et al. (2003)	
Al-Mubaid and Nguyen (2009)	
Pedersen et al. (2007)	
Sánchez et al. (2012)	
Hadj Taieb et al. (2014b)	

Table 11: Synset-valued data file based on the RG65 dataset that includes the similarity values reported by all the non IC-based similarity measures shown in table 1. Every file contains a set of taxonomical features and the degree of similarity between the two synsets corresponding to each synset pair between two input words.

Chapter 15

Reproducible Experiments based on ReproZip

This page intentionally left blank.

WordNet-based word similarity reproducible experiments based on HESML V1R1 and ReproZip

Published: 8 Sep 2016 | **Version 1** | DOI: 10.17632/65pxgskhz9.1

Contributor(s): [Juan J. Lastra-Díaz](#), [Ana Garcia-Serrano](#)

Description of this data

This dataset is provided as supplementary material of the paper by Lastra-Díaz, J. J., & García-Serrano, A. (2016). HESML: a scalable ontology-based semantic similarity measures library with a set of reproducible experiments and a replication dataset. *Information Systems*.

This dataset contains a ReproZip reproducible experiment file, called "HESMLv1r1_reproducible_exps.rpz", which allows the experimental surveys on word similarity on WordNet introduced in the three papers below to be reproduced exactly.

[1] Lastra-Díaz, J. J., & García-Serrano, A. (2015). A novel family of IC-based similarity measures with a detailed experimental survey on WordNet. *Engineering Applications of Artificial Intelligence Journal*, 46, 140–153. <http://dx.doi.org/10.1016/j.engappai.2015.09.006>

[2] Lastra-Díaz, J. J., & García-Serrano, A. (2015). A new family of information content models with an experimental survey on WordNet. *Knowledge-Based Systems*, 89, 509–526. <http://dx.doi.org/10.1016/j.knosys.2015.08.019>

[3] Lastra-Díaz, J. J., & García-Serrano, A. (2016). A refinement of the well-founded Information Content models with a very detailed experimental survey on WordNet (No. TR-2016-01). NLP and IR Research Group. ETSI Informática. Universidad Nacional de Educación a Distancia (UNED). <http://e-spacio.uned.es/fez/view/bibliuned:DptoLSI-ETSI-Informes-Jlastra-refinement>

Experiment data files

[Download all files \(2\)](#)

 RawOutputFiles.zip

49 KB

This file contains the raw output files generated by the reproducible experiments in the accompanying ReproZip file called "HESMLv1r1_reproducible_exps.rpz".

 HESMLv1r1_reproducible_exps.rpz

56 MB

This a ReproZip file which contains a set of reproducible experiments which allow the results reported in the three aforementioned papers by Lastra-Díaz and García-Serrano to be reproduced by using ReproUnzip.

Steps to reproduce

In order to reproduce the experiments contained in the HESMLv1r1_reproducible_exps.rpz file, you should follow the detailed instructions in the main companion paper below.

Lastra-Díaz, J. J., & García-Serrano, A. (2016). HESML: a scalable ontology-based semantic similarity measures library with a set of reproducible experiments and a replication dataset. *Information Systems*.

Related links

[A novel family of IC-based similarity measures with a detailed experimental survey on WordNet](#)

article is related to this dataset

[doi:10.1016/j.engappai.2015.09.006](https://doi.org/10.1016/j.engappai.2015.09.006)

[A new family of information content models with an experimental survey on WordNet](#)

article is related to this dataset

[doi:10.1016/j.knosys.2015.08.019](https://doi.org/10.1016/j.knosys.2015.08.019)

<http://e-spacio.uned.es/fez/view/bibliuned:DptoLSI-ETSI-Informes-Jlastra-refinement>

article is related to this dataset

<http://dx.doi.org/10.17632/t87s78dg78.1>

dataset is related to this dataset

Associated article

peer reviewed

This data is associated with the following peer reviewed publication:

[HESML: A scalable ontology-based semantic similarity measures library with a set of reproducible experiments and a replication dataset](#)

[Cite this article](#)

Published in:

Information Systems

Chapter 16

Benchmarks between Semantic Measure Libraries

This page intentionally left blank.

HESML_vs_SML: scalability and performance benchmarks between the HESML V1R2 and SML 0.9 semantic measures libraries

Published: 21 Dec 2016 | **Version 1** | DOI: 10.17632/5hg3z85wf4.1

Contributor(s): [Juan J. Lastra-Diaz](#), [Ana Garcia-Serrano](#)

Description of this data

This dataset introduces a companion reproducibility Java console program, called HESML_vs_SML_test.jar, of the work introduced by Lastra-Díaz and García-Serrano [1]. This latter work introduces the Half-Edge Semantic Measures Library (HESML), and carries-out an experimental survey between HESML V1R2, the Semantic Measures Library (SML) 0.9 [2] and the WNetSS [4] semantic measures libraries.

The HESML_vs_SML_test.jar program runs the set of performance and scalability benchmarks detailed in [1] and generates the figures and tables of results reported in the aforementioned work, which are also enclosed as complementary files of this dataset (see files below).

Licensing note:

The 'HESML_vs_SML_test.jar' program is based on the HESML V1R2 [3], SML 0.9 [2] and WNetSS [4] semantic measures libraries, and it includes these libraries in its distribution, as well as WordNet 3.0 [6] and the SimLex665 [5] dataset. Thus, if you use this dataset, you should also cite the works related to these resources.

References:


- [1] Lastra-Díaz, J. J., and García-Serrano, A. (2016). HESML: a scalable ontology-based semantic similarity measures library with a set of reproducible experiments and a replication dataset. To appear in Information Systems Journal.
- [2] Harispe, S., Ranwez, S., Janaqi, S., and Montmain, J. (2014). The Semantic Measures Library: Assessing Semantic Similarity from Knowledge Representation Analysis. In E. Métais, M. Roche, & M. Teisseire (Eds.), Proc. of the 19th International Conference on Applications of Natural Language to Information Systems (NLDB 2014) (Vol. 8455, pp. 254–257). Montpellier, France: Springer. http://dx.doi.org/10.1007/978-3-319-07983-7_37
- [3] Lastra-Díaz, J. J., & García-Serrano, A. (2016). HESML V1R2 Java software library of ontology-based semantic similarity measures and information content models. Mendeley Data, v2. <https://doi.org/10.17632/t87s78dg78.2>
- [4] Ben Aouicha, M., Taieb, M. A. H., and Ben Hamadou, A. (2016). SISR: System for integrating semantic relatedness and similarity measures. Soft Computing, 1–25. <http://dx.doi.org/10.1007/s00500-016-2438-x>
- [5] Hill, F., Reichart, R., & Korhonen, A. (2015). SimLex-999: Evaluating Semantic Models with (Genuine) Similarity Estimation. Computational Linguistics, 41(4), 665–695.

http://dx.doi.org/10.1162/COLI_a_00237

[6] Miller, G. A. (1995). WordNet: A Lexical Database for English. *Communications of the ACM*, 38(11), 39–41. <http://dx.doi.org/10.1145/219717.219748>

Experiment data files

[Download all files \(5\)](#)

 benchmarks_HESML_vs_SML.csv


3 KB

Raw output file containing the results of the benchmarks. This file was generated by HESML_vs_SML_test as output file on a Windows 10 workstation with 8 Gb RAM and an Intel Core i7-5500U CPU @ 2.40 GHz.

 HESML_vs_SML.pdf

5 KB

This figure shows the results of the main scalability benchmarks between HESML and SML. The running times are averaged as regards the size of the taxonomy or the number of topological queries used in each benchmark.

 final_results_SimLex665.csv


0.23 KB

Raw output file containing the running times reported by the HESML V1R2, SML 0.9 and WNetSS semantic measures libraries in the evaluation of the Jiang-Conrath similarity measure with the Seco et al. IC model in the SimLex665 dataset.

 IS_HESML_figure3_and_table18.r

6 KB

R script file with the aim of reproducing the figure shown in the HESML_vs_SML.pdf file above from the benchmarks_HESML_vs_SML.csv file above.

 HESML_vs_SML_test.zip

42 MB

Java source files, executable and NetBeans project of the HESML_vs_SML_test program which runs a set of benchmarks between the HESML V1R2, SML 0.9 and WNetSS semantic measures libraries.

Steps to reproduce

System requirements: a Java8-compliant workstation with at least 8 Gb RAM.

The HESML_vs_SML_test.zip file contains the source files and compiled versions of the HESML_vs_SML_test.jar and all the aforementioned semantic measures libraries, thus, you only need to run the program. However, in order to compile HESML_vs_SML_test from its source files, you need to install NetBeans 8.0 or higher and the Java SDK 8.0.

Running of the benchmarks:

The first group of benchmarks evaluates the running-time and caching ratio in a side-by-side comparison between the most significant topological algorithms implemented by HESML and SML.

(1) Download the HESML_VS_SML_test.zip file above and extract it onto your hard drive, then follow the steps 2-4 below:

(2) Open a Linux or Windows command console in the main HESML_VS_SML_test directory and run the following command:

```
$prompt:> java -Xms4096m -Xmx4096m -jar dist\HESML_VS_SML_test.jar <output_results.csv>
```

(3) Import the raw output file with LibreOffice or MS-Excel to obtain the data as shown in benchmarks_HESML_vs_SML.csv file above

(4) Install and open the R statistics package, then follow the following steps: (a) select the "File->Open script" menu and load the 'IS_HESML_figure3_and_table18.r' script file above; (b) edit the first two lines of the script code in order to set the path of the input directory and the input 'output_results.csv' file generated in the step 2 above; and finally, (c) select the "Edit->Run all" menu in order to generate the figure in the HESML_vs_SML.pdf file above.

The output csv file obtained in step 2 above will be identical to the complementary 'benchmarks_HESML_vs_SML.csv' file. However, it will show the running times on your experimental platform.

The second benchmark evaluates the running time of HESML, SML and WNetSS in the evaluation of the Jiang-Conrath similarity measure with the Seco et al. IC model in the SimLex665 dataset. In order to reproduce the WordNet-based similarity benchmark reported in table 19 of [1] and the 'final_results-SimLex665.csv' file above, you should follow the steps 5-8 below:

(5) Install MySQL community edition in your workstation (demanded by WNetSS).

(6) Open a Linux or Windows command console in the HESML_VS_SML_test directory and run the command below, which carries out the off-line pre-processing tasks of WNetSS in order to load WordNet 3.0 and all its topological information in the MySQL server. This task could take a few hours in a modern workstation.

```
$prompt:> java -Xms4096m -Xmx4096m -jar dist\HESML_VS_SML_test.jar -WNetSS_Setup mySqlRootPassword
```

(7) From the same Linux or Windows command console run the following command:

```
$prompt:> java -Xms4096m -Xmx4096m -jar dist\HESML_VS_SML_test.jar -WNetSS mySqlRootPassword <output_results.csv>
```

(8) Import the output file with LibreOffice or MS-Excel to obtain the data shown in the final_results_SimLex665.csv file above.

Related links

<http://dx.doi.org/10.17632/t87s78dg78.2>

software is source of this dataset

[The Semantic Measures Library: Assessing Semantic Similarity from Knowledge Representation Analysis](#)

