Web Spam Detection: New Classification Features Based on Qualified Link Analysis and Language Models

Lourdes Araujo and Juan Martinez-Romo

***Abstract*—Web spam is a serious problem for search engines because the quality of their results can be severely degraded by the presence of this kind of page. In this paper, we present an efficient spam detection system based on a classifier that combines new link-based features with language-model (LM)-based ones. These features are not only related to quantitative data extracted from the Web pages, but also to qualitative properties, mainly of the page links. We consider, for instance, the ability of a search engine to find, using information provided by the page for a given link, the page that the link actually points at. This can be regarded as indicative of the link reliability. We also check the coherence between a page and another one pointed at by any of its links. Two pages linked by a hyperlink should be semantically related, by at least a weak contextual relation. Thus, we apply an LM approach to different sources of information from a Web page that belongs to the context of a link, in order to provide high-quality indicators of Web spam. We have specifically applied the Kullback–Leibler divergence on different combinations of these sources of information in order to characterize the relationship between two linked pages. The result is a system that significantly improves the detection of Web spam using fewer features, on two large and public datasets such as `WEBSPAM-UK2006` and `WEBSPAM-UK2007`.**

***Index Terms*—Content analysis, information retrieval, language models (LMs), link integrity, Web spam detection.**

I. Introduction

WEB spam is one of the main current problems of search engines because it strongly degrades the quality of the results. Many people become frustrated by constantly finding spam sites when they look for legitimate content. In addition, Web spam has an economic impact since a high ranking provides large free advertising and so an increase in the Web traffic volume. During recent years, there have been many advances in the detection of these fraudulent pages but, in response, new spam techniques have appeared. Research in this area has became an arms race to fight an adversary who constantly uses more and more sophisticated methods. For this reason, it is necessary to improve anti-spam techniques to get over these attacks. Web spam, or spamdexing, includes all techniques used for the purpose of getting an undeservedly high rank. In general terms, there are three types of Web spam: link spam, content spam, and cloaking, a technique in which the content presented to the search engine spider is different to that presented to the browser of the user. However, link and content spam are the most common types, and the ones considered in this work. According to Davison [10], link spam can be defined as "links between pages that are present for reasons other than merit." Link spam consists of the creation of a link structure to take advantage of link-based ranking algorithms, such as PageRank, which gives a higher ranking to a website the more other highly ranked websites link to it. Content spam includes all techniques that involve altering the logical view that a search engine has over the page contents [13], for instance, by inserting keywords that are more related to popular query terms than to the actual content of the page.

One of the most successful techniques for Web spam detection, as it can be seen in the AIRWeb competition,[1] is the definition of features which take different values for spam and nonspam pages. These features are thus used to implement a classifier able to detect spam pages.

In this work, we also adopt this scheme and propose new features to characterize Web spam pages. However, while most previous works using content and link-based features to detect spam are focused on quantitative features, in this work, we propose several new qualitative features [21] to improve Web spam detection grouped in two sets: 1) a group of link-based features which check the reliability of links, and 2) a group of content-based features extracted with the help of a language-model (LM) approach. Finally, we build an automatic classifier that combines both types of features, reaching a precision that improves the results of each type separately and those obtained by other proposals.

Some of the considered features are related to the quality of the links in the page. Typically, links in nonspam pages point as well to nonspam pages and are appropriately described by the corresponding anchor text and its context. Accordingly, we propose a number of features which capture these differences with the spam pages. Some of these features are based on the behavior of standard search engines, applied to queries composed of pieces of information that pages provide for their links. It is natural that the information associated to a link—terms in the URL, in the anchor text and other terms in the context of the link—allows recovery in a relevant position the page actually pointed at by the link in nonspam pages. So, we expect a dif-

Manuscript received August 25, 2009; revised March 18, 2010; accepted May 05, 2010. Date of publication June 01, 2010; date of current version August 13, 2010. This work was supported in part by the Spanish Ministry of Science and Innovation within the project QEAVis-Catiex (TIN2007-67581-C02-01) and in part by the Regional Government of Madrid under the Research Network MA2VICMR (S2009/TIC-1542). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Darko Kirovski.

The authors are with the NLP & IR Group, UNED, Madrid 28040, Spain (e-mail: lurdes@lsi.uned.es; juaner@lsi.uned.es).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIFS.2010.2050767

[1] Available: http://airweb.cse.lehigh.edu/

ferent behavior in the recovering capacity for reliable and deceptive links, which can be used as a distinction feature for our classifier. Thus, we use the whole set of techniques and sophisticated refinement provided by standard search engines. By using a standard search engine, such as Yahoo!, we are taking advantage of this technology, as a black box, to recover on the top relevant pages concerning some selected terms. It does not matter for our system if this technology changes, as long as its purpose continues to be retrieving the most relevant pages for a query at the first positions. We have also introduced other features related to the links, such as the existence of broken links in the page and the presence of links pointing to spam pages.

Other sets of features considered in this work try to capture the coherence between a page and the pages it points to. In general, some degree of relationship is expected between the information associated to a link in the studied page and the content of the pointed page. To measure this coherence, we resort to an LM approach. LMs [20] are probabilistic methods which have been developed to capture linguistic features hidden in texts, such as the probability of words or word sequences in a language. LMs have been successfully used in speech recognition, machine translation, part-of-speech tagging, parsing, and information retrieval. Previous works have proved that LM disagreement techniques are very efficient in tasks such as blocking blog spam [18] or detecting nepotistic links [3]. Thus, we use an extension of the basic language modeling approach to analyze several sources of information extracted from each website in the collection. Certainly spammers can try to take advantage of some of our techniques, but several of the proposed features try to capture the coherence between pages, and others the quality of their links. Thus, we believe that spammers could use mechanisms not to be detected as spam by some of our features, but it is unlikely that they are able to maintain appropriate rates for all the features at the same time.

A preliminary study of this kind of feature related to content coherence [17] has revealed its usefulness in spam detection. The set of features used in this preliminary work have now been extended with new features related to sources of information associated to each link (the URL, the surrounding anchor text, the title page, or the meta tags), as well as to the page pointed at (content of the page, title, or meta tags). We make an LM from every source of information, and then calculate the Kullback–Leibler (KL) divergence [8] between their respective LMs. The coherence between different combinations of the considered sources defines a set of features for our spam classifier. These features are combined in the current work with the set of features related to the link quality, improving substantially the results of the preliminary work. For that, we have studied different quality parameters of a website extracted from the contextual information of their links. Another contribution of this work is the adaptation of a complex system [15], [16] for the recovery of broken Web links to the detection of Web spam.

In order to evaluate the system, it is necessary to resort to some collection of pages that have already been manually labeled as *spam* or *nonspam*, and thus can be used to train and evaluate the system. We have evaluated the capacity of our system to detect the two major types of Web spam, link and content, in two widely used Web spam labeled collections [6], containing pages from these two types. Results have revealed that each type of considered feature does not achieve an important improvement over the results obtained with the standard features provided in the evaluation collection. However, when all the different types of features are included in the classifier, the system achieves a very high improvement over previous results.

The remainder of the paper is as follows: Section II presents previous work in the Web spam research area; Section III describes the qualified-link (QL) analysis; Section IV studies the suitability of different sources of information to provide features based on divergence measures; and Section V is devoted to describing the methodology adopted for evaluation and the experiments proposed, as well as the results obtained on two public datasets. Finally, Section VI draws the main conclusions.

II. Background and Related Work

There are works in the literature devoted to the different kind of spam considered in this work. Some of the highlights of the case of link spam are: Becchetti *et al.* [2], who used automatic classifiers to detect link-based spam and Benczúr *et al.* [4], who analyzed supporting sets and PageRank contributions for building an algorithm to detect link spam. Other works are focused on content spam: Ntoulas *et al.* [19] introduced new features based on checksums and word weighting techniques and Webb *et al.* [22], who proposed a real-time system for web spam classification by using HTTP response headers to extract several features. Studies also exist that have combined the detection of different types of spam: Abernethy *et al.* [1] trained a support vector machine (SVM) classifier with content and link data and Castillo *et al.* [7] combined content and topology information in a cost-sensitive tree.

Closest to our research are the works by Mishne *et al.* [18], that apply LMs to Blog spam detection. Here, the authors estimate LMs from the original post and each comment in a Blog and then they compare these models using a variation on the Interpolated Aggregate Smoothing. In particular, this measure calculates the smoothed KL divergence between the LM of a short fragment of text (original post) and a combined LM of knowledge preceding this text (previous comments). They collected 50 random blog posts, along with the 1024 comments posted to them and although they did not get very good results, they propose a model expansion that should improve the performance. Benczúr *et al.* [3] proposed to detect nepotistic links using LMs tested on a 31M page crawl of the .de domain with a manually classified 1000-page random sample. In this method, a link is down-weighted if the LMs from its source and target page have a great disagreement. Specifically, they used KL divergence between the unigram LM of the target and source pages. Then, they fed suspicious edges into a weighted PageRank calculation to obtain NRank, the "nepotism rank" of the pointed page, which was subtracted from the original PageRank value. We share, with this approach, the assumption that pages that are connected by non-nepotistic links must be sufficiently similar. Qi *et al.* [21] distinguished between QLs and advertising or spam, using six similarity measures considering the issue of computational complexity: Host, URL, Topic Vector, TF-IDF content, Anchor Text, and Nonanchor Text. To calculate these measures they used methods such as Cosine, Dice, or Naive Bayes over the URL terms, anchor texts, or content. They also compared this method with Hits and PageRank ranking approaches, introducing two measures: Qualified HITS and Qualified PageRank. Through experiments on 53 query-specific datasets, they showed that their approach improved precision by 9% compared to the Bharat and Henzinger [5] HITS variation proposal.

III. Qualified Link Analysis

We propose a deep analysis of Web links from the standpoint of quality as defined in [21]. This qualitative analysis has been designed to study neither the network topology, nor link characteristics in a graph. With this sort of analysis, we mainly try to find nepotistic links [10], [3] that are present for reasons other than merit. For that, we have studied different quality parameters from a website. Some of them concern the page links, such as testing if they are broken or measuring the difference between internal and external links or between outgoing and incoming links. Others refer to the content of the anchor text: whether it is just a URL, a number, a punctuation mark, or even just an empty chain. Finally, others are related to different aspects of the coherence between a link (its anchor text, surrounding words, etc.) and the pointed page, and between the page containing the link and the pointed page.

For this task, we have developed an information retrieval system which provides us with a quality factor from every page which is represented by a set of features about its links. This information is very useful because we will be able to detect a large number of links whose sole purpose is to move up in the ranking of a search engine through building a network of link farms.

A time gap between the labeling process of the reference collection and the moment in which these features are extracted using the recovery system could be noticeable. Between these two time frames, certain features from the pages pointed by the pages in the collection could have changed. From our standpoint, this time gap could only have worsened the results in spam detection, so the results that are shown could be improved if all features and the labeling process were obtained at the same time.

A. Analyzing Web Links

The information retrieval system analyzes the links in a page and extracts several features from that page. The system not only offers information about the number of links whose pointed page can be recovered using information from the link and the page that contains it, but also data about every link. This system is based on classical information retrieval techniques and natural language processing, and it mainly consists of two stages.

1) **Extraction of relevant information on a link.** There are many works which have analyzed the importance of the anchor text as a source of information [9], thus we use the anchor text as the main source of information to recover a link. However, there are many cases in which the anchor text does not contain enough information. For this reason, the system performs a terminology extraction from other sources of information such as the URL, the page that contains the link, the context of the anchor text, and a cached page version of the analyzed link that can be stored in a search engine (Yahoo!) or digital library (Wayback Machine). The system uses several terminology extraction approaches based on frequency [term frequency-inverse document frequency (TF-IDF[2])] and statistical language modeling (specifically KL divergence), depending on the source of information considered. Specifically, TF-IDF is used to extract terminology from the page that contains the link, which could be unrelated to the searched page, and KL divergence to extract terminology from the cache page, which, if it exists, contains relevant terminology for sure.
2) **Construction of complex queries and request to a search engine.** The original query is composed of the terms extracted from the anchor text, and this query is expanded using the terms extracted from the other sources of information considered. The different expanded queries are submitted to the selected search engine (Yahoo!), and the top ten ranked documents are retrieved. In this paper, we consider that a link has been recovered if the page pointed by the link is in the set of pages retrieved with some of the queries.

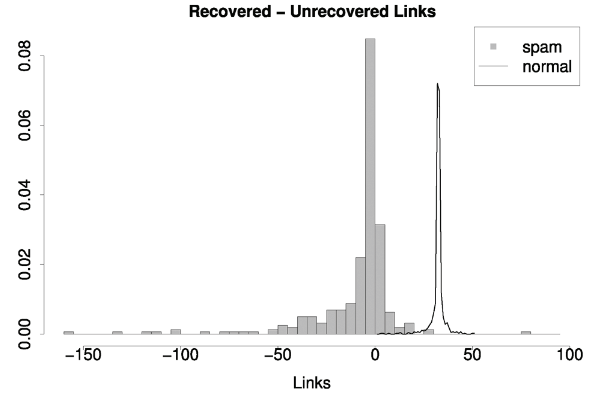

Fig. 1. Histogram of distributions of the difference between recovered and nonrecovered links. The horizontal axis corresponds to the number of recovered links minus the number of nonrecovered. The vertical axis corresponds to density.

Six measures are extracted in this work and described below. We have considered different features for each measure.

Recovery Degree: The most important feature that is extracted thanks to the recovery system is precisely the degree of recovered links. For every page the system tries to retrieve all their links and as result, three values are obtained: 1) the number of recovered links (retrieved within the top ten results of the search), 2) the number of not recovered links, and 3) the difference between both previous values, which is represented in Fig. 1. In the figure, we can observe that the spam pages concentrate on a separate area of the distribution, which allows us to distinguish them. We can also observe than the rate of recovered links with respect to nonrecovered is clearly higher in the nonspam pages, thus providing a very useful feature for the classifier. The degree of recovered links can be understood as a coherence measure between the analyzed page, one of its links, and the page pointed by this link. The intuition in the interpretation of this feature is that a page that belongs to a link

[2] TF-IDF is a weight used to evaluate the relevance of a term of a document in a collection or corpus. This relevance is proportional to the number of times the term appears in the document and inversely proportional to the number of documents containing the term in the corpus.

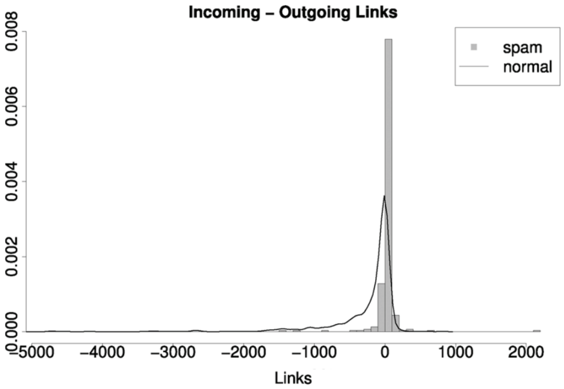

Fig. 2. Histogram of distributions of the difference between incoming and outgoing links.

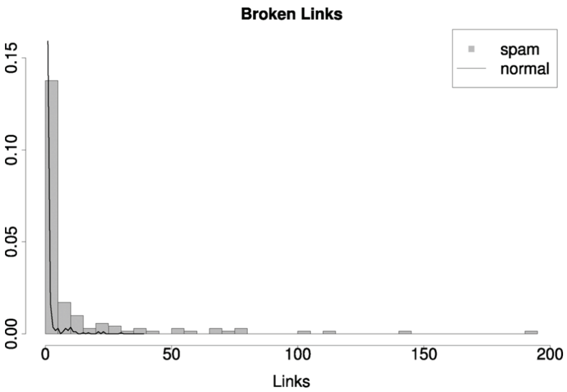

Fig. 4. Histogram of the number of broken links.

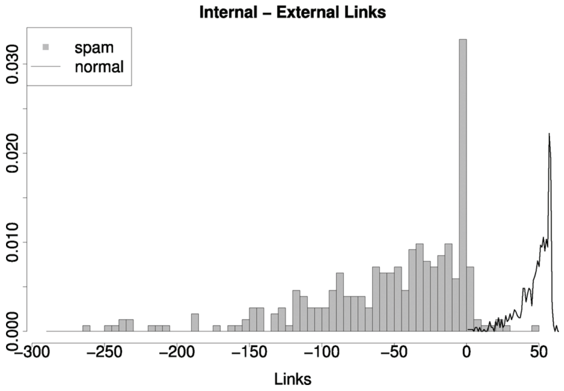

Fig. 3. Histogram of distributions of the difference between internal and external links.

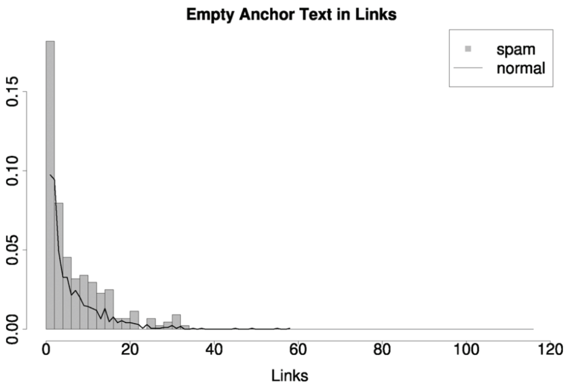

Fig. 5. Histogram of distributions of the number of links whose anchor text is empty.

farm is linking to other unknown pages with the only purpose of appearing in the top of the ranking of search engines. Therefore, these links are difficult to retrieve. Thus, the more negative the difference between the recovered and not recovered links, the greater the likelihood that this site is applying spam techniques.

Incoming–Outgoing: It is well-known that spam pages link to nonspam pages, but nonspam pages do not link to spam pages. Taking advantage of the possibilities of the system to submit queries to a search engine, we have included a new query to request to the search engine, how many sites point to the analyzed page (incoming links). Fig. 2 represents the difference in the amount of links of each type. We can observe the difference shape of the graphic for spam and nonspam pages. In addition, we have used the number of outgoing links as another feature.

External–Internal: Several theories exist about the impact of internal and external links in the PageRank of a site. Although there is no definitive evidence to prove it, we think that many websites apply these theories. For this reason, we have taken the number of external and internal links as features. Fig. 3 represents the rate of these two types of links for spam and nonspam pages, showing that this feature takes negative values for spam pages, and positive for nonspam pages.

Broken Links: Broken links are a common problem for both spam and nonspam pages, even when this sort of link has a negative impact in the PageRank. Fig. 4 compares this feature for spam and nonspam pages. The number of spam pages is higher in almost the whole range of numbers of the broken links considered.

Anchor Text Typology: It is usual that spam pages contain text and links automatically generated. Moreover, the anchor text of many links are usually generated thinking in the context of the search engines instead of the users. Thus, we have selected four features in order to measure the number of links that are formed only by 1) punctuation marks, 2) digits, 3) a URL, and 4) an empty chain. Fig. 5 shows the histogram for those links whose anchor text is empty, for spam and nonspam pages. We can see that the shape is different in both cases. Though there are areas where the values overlap for spam and nonspam pages, we have to take into account that the classifier uses a whole set of features, by assigning different weights to the most appropriates in every case.

Thus, we have in total 12 features for each Web page. Analyzing the histograms from each feature, we can conclude that the features that offer the best divergence among the spam and nonspam pages are the following (in relevance order): 1) difference between recovered and not recovered links, 2) number of links with an empty anchor text, and 3) difference between external and internal links. However, all the features contribute to the performance of the classifier because each of them can discriminate different cases.

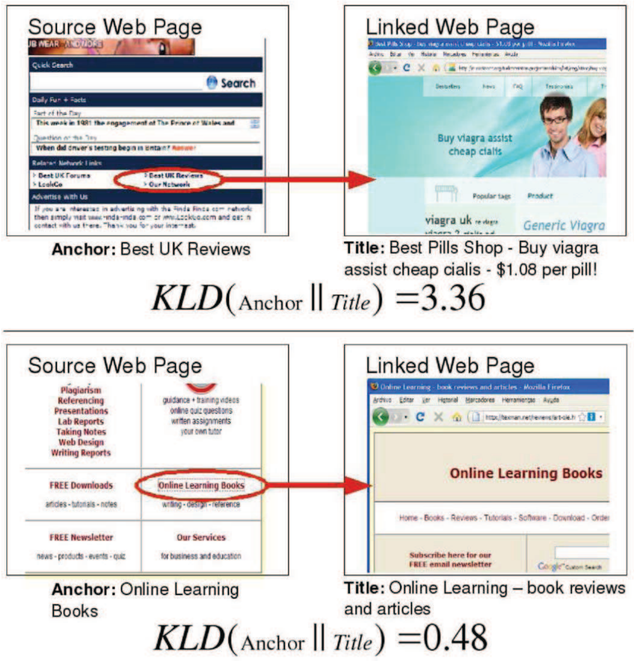
Source Web Page
Linked Web Page
Anchor: Best UK Reviews
Title: Best Pills Shop - Buy viagra assist cheap cialis - $1.08 per pill!
KLD(Anchor || Title) =3.36
Source Web Page
Linked Web Page
Online Learning Books
Anchor: Online Learning Books
Title: Online Learning – book reviews and articles
KLD(Anchor || Title) =0.48

Fig. 6. Examples (spam and nonspam) of KL divergence applied to the anchor text of a link and the title of the page pointed by this link.

IV. Language Models

One of the most successful methods based on term distribution analysis uses the concept of KL divergence [8] to compute the divergence between the probability distributions of terms of two particular documents considered.

We have applied KL divergence to measure the differences between two text units of the source and target pages. Specifically, we look at the differences in the term distribution between two text units by computing the KL divergence

$$\mathrm{KLD}(T_1 \| T_2) = \sum_{t \in T_1} P_{T_1}(t) \log \frac{P_{T_1}(t)}{P_{T_2}(t)} \tag{1}$$

where $P_{T_1}(t)$ is the probability of the term t in the first text unit, and $P_{T_2}(t)$ is the probability of the term t in the second text unit.

The LMs that we use estimate maximum likelihood of the unigram occurrence probabilities.

A. LM-Based Features

We characterize the relationship between two linked Web pages according to different values of divergence. These values are obtained by calculating the KL divergence between one or more sources of information from each page. In Fig. 6, two examples are shown, illustrating the KL divergence applied to the anchor text of a link and the title of the page pointed by this link.

In particular, we consider the following three sources of information from the source page:

Anchor Text: When a page links to another, this page has only a way to convince a user to visit this link, that is by showing relevant and summarized information of the target page. This is the function of the anchor text. Therefore, a great divergence between this piece of text and the linked page shows a clear evidence of spam. In addition, Mishne *et al.* [18] and Benczúr *et al.* [3] proved that disagreement between anchor text and the target content is a very useful measure to detect spam.

Surrounding Anchor Text: Sometimes anchor terms provide little or no descriptive value. Let us imagine a link whose anchor text is "click here." For this reason, text surrounding a link can provide contextual information about the pointed page. Moreover, in [3], a better behavior is observed when the anchor text is extended with neighboring words. In our experiments, we have used several words around the anchor text (seven per side) to extend it, though we took into account HTML block-level elements and punctuation marks.

URL Terms: Besides the anchor text, the only information available of a link is its URL. A URL is mainly composed of a protocol, a domain, a path, and a file. These elements are composed of terms that can provide rich information from the target page. During recent years, because of the increasing use of search engines, search engine optimization (SEO) techniques exist that try to exploit the importance of URL terms in a request. Thus, if we have a URL such as "www.domain.com/viagra-youtube-free-download-poker-online.html", and after visiting this page, a pornographic site, it could be said that this page uses spam techniques. Therefore, we have retrieved the most relevant terms from a URL in order to calculate the divergence with the content of the target page. To extract these most relevant terms, first of all, we have built an LM with terms from URLs in the Open Directory Project (ODP) public list. Afterwards, with help of this collection of URLs, we have applied the KL divergence in order to know the most relevant terms in a certain URL. Finally, we use the top 60% of these terms because this value has provided the best results in some preliminary experiments.

We also get the following three sources of information from the target page:

Title: Jin *et al.* [14] observed that document titles bear a close resemblance to queries, and that they are produced by a similar mental process. Eiron *et al.* [11] studied the similarity of title and anchor text and they concluded that both titles and anchor text capture some notion of what a document is about, though these sources of information are linguistically dissimilar. In addition, it is well-known that anchor text, terms of a URL, and terms of the Web page title, have a great impact when search engines decide whether a page is relevant to a query. In other words, spammers perform engineering tasks in order to set key terms in these sources of information. Therefore, divergence between these sources of information, from source and target pages, reports a great usefulness in the detection of Web spam.

Page content: The page content is the main source of information that is usually available. Although in many cases, the title and meta tags from the target page are not available, most Web pages have at least a certain amount of text. Previous works that have studied the relationship between two linked Web pages, have usually considered the content of the target page in order to extract any data and/or measure. Qi *et al.* [21] used the TF-IDF content similarity of two Web pages by measuring the term-based similarity among their 1) textual content, 2) anchor text, and 3) nonanchor text. In addition, Mishne *et al.* [18] compared two LMs between blog posts and pages linked by comments, and Benczúr [3] *et al.* proved that disagreement

TABLE I
COMBINATION OF DIFFERENT SOURCES OF INFORMATION USED TO CALCULATE THE KL DIVERGENCE

Combination of different Sources of Information
Page Content (P)
Anchor Text $(A \rightarrow P)$
Surrounding Anchor Text $(S \rightarrow P)$
URL Terms $(U \rightarrow P)$
Anchor Text ∪ URL Terms $(AU \rightarrow P)$
Surrounding Anchor Text ∪ URL Terms $(SU \rightarrow P)$
Title vs Page $(T \rightarrow P)$
Meta Tags vs Page $(M \rightarrow P)$
Title (T)
Anchor Text $(A \rightarrow T)$
Surrounding Anchor Text $(S \rightarrow T)$
URL Terms $(U \rightarrow T)$
Surrounding Anchor Text ∪ URL Terms $(SU \rightarrow T)$
Meta Tags (M)
Anchor Text $(A \rightarrow M)$
Surrounding Anchor Text $(S \rightarrow M)$
Surrounding Anchor Text ∪ URL Terms $(SU \rightarrow M)$

between anchor text and the target content is a very useful measure to detect spam.

Meta Tags: Meta tags provide structured meta data about a Web page and they are used in SEO. Although they have been the target of spammers for a long time and search engines consider these data less and less, there are pages still using them because of their clear usefulness. In particular we have considered the attributes "description" and "keywords" from meta tags to build a virtual document with their terms. We have decided to use these data to calculate its divergence with other sources of information from the source page, such as anchor text and surrounding anchor text, and from the target page such as page content and URL terms. Although meta tags are only found at between 30%–40% of the sites, when they are located in a Web page, their usefulness is very high.

Many combinations of these sources of information could be used to measure the divergence between two Web pages. However, considering the issue of computational complexity, he have chosen a set of features that are easy to compute and that are useful in Web spam detection. Moreover, we have used Lucene [12] to carry out the computation, which is a source information retrieval library. These features are described below.

B. Combination of Sources of Information

In addition to using these sources of information individually, we have combined some of them from the source page with the goal of creating virtual documents which provide richer information. As we have seen above, we have used Anchor Text (A), Surrounding Anchor Text (S), and URL terms (U) as sources of information. We also propose to create two new sources of information: 1) combining Anchor Text and URL terms (AU) and 2) combining Surrounding Anchor Text and URL terms (SU). In addition, we have considered other sources of information from the target page: Content Page (P), Title (T), and Meta Tags (M). We have also ruled out the use of any combination due to the limited relationship between these sources of information. Table I summarizes all 14 features used in this work. The group on the top corresponds to divergences between different data (or combinations of them) in the source page and the pointed page (P). The group in the middle corresponds to divergences between data in the source page and the title of the pointed page.

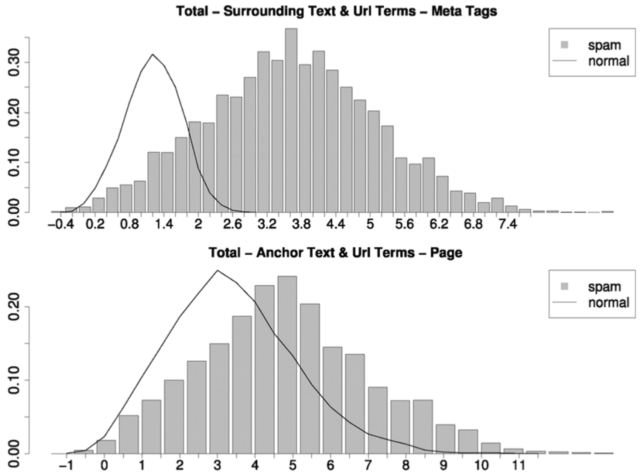

Fig. 7. (Top) Histogram of KL divergence distribution between a combination of Anchor Text, Surrounding Anchor Text (which also includes the Anchor Text), and URL Terms from the source page and the target Page Title. (Bottom) Histogram of KL divergence between a combination of Anchor Text and URL Terms from the source page and the target Page Content. The mark Total refers to using both, internal and external links, to compute these data.

And the last group corresponds to divergence between data in the source page and meta tags associated to the pointed page.

In many cases, we can find anchors with a small number of terms that sometimes mislead our results. However, by combining different sources of information such as Anchor text, Surrounding Anchor text, and URL terms, we can obtain a more descriptive language. Furthermore, despite the fact that single sources of information offer interesting divergence values between spam and nonspam pages, the best measures proposed in this work are obtained by combining different sources of information. As it can be seen in Fig. 7, these combinations of sources of information get a high divergence between spam and nonspam pages. First of all, we can observe the different shapes of the distribution for spam and nonspam pages. Both distributions have Gaussian shapes, but nonspam histograms are more compact and their means are near KL ≈ 1.2 and KL ≈ 3.5, respectively. On the other hand, spam histograms are wider, and their means are near KL ≈ 4 and KL ≈ 5, respectively. We can also observe in the top figure, corresponding to the divergence between the terms from the text surrounding the link, including the anchor text, plus URL terms from the source page and meta tags associated to the pointed page, that there is a range of divergence values for which this feature can effectively discriminate spam and nonspam pages (for values greater than 2) with a high degree of confidence. Results shown in the bottom figure, corresponding to the divergence between terms from the anchor text plus URL terms from the source page and the pointed page content, also show a range of values (for values greater than 6), although it is smaller in this case, for which we can discriminate spam and nonspam pages.

C. Internal and External Links

SEO Websites and Blogs[3] have published some articles which assert that the relationship between internal and external links, i.e., a ratio between the number of such links, is important to

[3] Available: www.seo-theory.com

obtain a higher PageRank. Thus, internal and external links in a page would have impact on the ranking provided by a search engine. This suggests that spammers may be using algorithms that take into account this information to promote their pages.

For these reasons, we have decided to distinguish internal and external links in order to carry out the divergence analysis. Therefore, for each Web page we have triple-features: 14 features for internal links, 14 features for external links, and 14 features for both internal and external links.

V. Experiments

We use two publicly available Web spam collections [6] based on crawls of the .uk Web domain done in May 2006 and May 2007, respectively. `WEBSPAM-UK2006` includes 77.9 million pages and over 3 billion links about 11 400 domains. `WEBSPAM-UK2007` include 105.9 million pages and over 3.7 billion links about 114 529 domains. These reference collections are tagged by a group of volunteers labeling domains as “nonspam,” “spam,” or “borderline.” In our experiments, we restricted the datasets using only domains labeled at least by two persons independently, and for which all assessors agreed. Moreover, ODP labels [6] are not taken into account.

After the described filtering steps, the `WEBSPAM-UK2006` dataset used in our experiments has 3083 domains, 1811 of which are labeled as “nonspam” and 1272 as “spam.” Moreover, nonspam domains have a mean of external and internal links of 12.1 and 30.6, respectively, and spam domains have a mean of external and internal links of 7.2 and 15.3. The `WEBSPAM-UK2007` dataset has 4166 domains, 4012 of which are labeled as “normal” and 154 as “spam.” In this dataset, normal domains have a mean of external and internal links of 3.7 and 13.4, respectively, and spam domains have a mean of external and internal links of 9.3 and 12.06.

The datasets are labeled at the domain level [6], so we have to aggregate all LM and QL features at this level. In order to carry out the divergence and link analysis and considering the computational cost, we only analyze one page per domain. Specifically, we select the homepage from the source page and every page pointed by any link in the source page. Furthermore, domains that have no outgoing links are discarded, so the final size of the dataset is slightly reduced.

The divergence analysis requires selecting links for each analyzed domain. We only analyze links that have some terms in the anchor text. Therefore, we filter out images, links to the same page (named anchor inside an HTML document), numbers, URLs, and empty strings. We also rule out links whose protocol is not HTTP and links to non-HTML documents. Finally, we obtain 42 divergence measures for each link in a Web page, and we estimate the mean of all links for each measure. A website is, therefore, represented by 42 features (mean values).

In the case of the QL analysis, we have used an adapted recovery system and we represent each domain with 12 features which have also been previously introduced.

The current average time for processing a web page is about 4 s. This time has a large variability (a standard deviation of 2 s) because it depends on the number of links of the page and the analyzed pages for every link.

We now present several experiments in identifying Web spam on the above described datasets.

A. Classification Algorithms

For the classification tasks, we have used the Weka [23] software because it contains a whole collection of machine-learning algorithms for data-mining tasks.

The first step to obtain the best results in the classification task is to select the most appropriate classifier. We selected different classification algorithms to evaluate the introduced features. In particular, we have chosen the following classification algorithms: Metacost, a cost-sensitive wrapper algorithm that takes the base classifier decision tree with bagging C4.5 [23]; Naive Bayes, a statistical classifier based on the Bayes’ theorem using the joint probabilities of sample observations to estimate the conditional probabilities of classes given an observation; Logistic Regression, a generalized linear model to apply regression to categorical variables; and finally, SVMs which aim at searching for a hyperplane that separates two classes of data with the largest margin.

We performed an extensive evaluation with these classifiers that are implemented in the Weka toolkit. We used the Weka J48 implementation of a decision tree, the Naïve Bayes and Logistic Regression algorithms, and the sequential minimal optimization (SMO) implementation of an SVM Polynomial kernel [23]. We used the default options of these algorithms, except in the case of the decision tree for which we set a reduced error pruning method. The algorithmic details of these classifiers are beyond the scope of this paper. Additional information about the classifiers is available in most standard machine-learning texts [23]. Optimizing the algorithm parameters could slightly improve our results, which can, therefore, be considered a lower bound of the performance we could obtain with our approach.

The evaluation of the learning schemes used in all the predictions of this paper was performed by a ten-fold cross-validation. For each evaluation, the dataset is split into ten equal partitions and is trained ten times. Every time, the classifier trains with nine out of the ten partitions and uses the tenth partition as test data. We have adopted a set of well-known [7] performance measures in Web spam research: true positive (TP or recall), false positive (FP) rate, and F-measure. F-measure combines precision P and recall R by $F = 2(PR)/(P + R)$. For evaluating the classification algorithms, we focus on the F-measure as it is a standard measure to summarize both precision P and recall R.

Table II shows the F-measure for all algorithms, based on the features we introduced in previous sections. The best classifier in most of the feature sets is the decision tree, followed by the SVM classifier. Even though the decision tree algorithm obtains the best results, over half of the spammers are classified as nonspammers. Section V-B, therefore, introduces costs for misclassifying spammers.

B. Cost-Sensitive Classifier

In the web spam collections that we use, the nonspam instances outnumber the spam ones to such an extent that the classifier accuracy improves by misclassifying a disproportionate number of spam instances. Thus, we think that errors for misclassifying nonspam pages as spam do not have the same impact that misclassifying a spam page as nonspam. We have used the Metacost [23] algorithm (cost-sensitive decision tree with bagging) implemented in Weka for classification, which allows

TABLE II
F-MEASURE FOR DECISION TREE, NAÏVE BAYES (NB), SVMs, AND LOGISTIC REGRESSION (LR) ALGORITHMS, BASED ON THE LMs AND QLs FEATURES FROM UK-2006 AND UK-2007 DATASETS

	F-Measure			
	UK-2006		UK-2007	
Alg.	**LM**	**QL**	**LM**	**QL**
C4.5	**0.55**	**0.67**	**0.24**	**0.32**
NB	**0.55**	0.57	0.20	0.18
SVM	0.54	0.65	0.22	**0.32**
LR	0.53	0.58	0.18	0.26

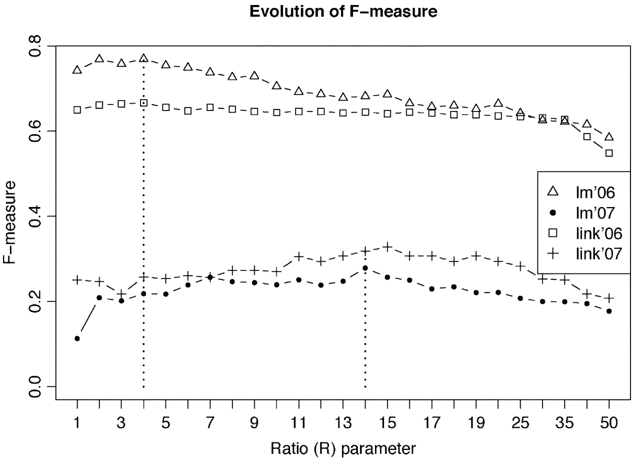

Fig. 8. Evolution of F-measure obtained by applying different costs to R in the confusion matrix. Content and Links-based features are used on WEBSPAM-UK2006 and WEBSPAM-UK2007.

TABLE III
FEATURES, TRUE POSITIVE (TP) RATE, FALSE POSITIVE (TP) RATE, F-MEASURE (F), AND AREA UNDER ROC CURVE (AUC) FOR WEB SPAM CLASSIFIERS USING DIFFERENT FEATURE SETS ON DATASET. THE BEST SCORES ARE MARKED IN BOLD

WEBSPAM-UK2006					
Feature Set	Features	*TP*	*FP*	*F*	*AUC*
Content (C)	98	0.61	0.08	0.63	0.82
Link (L)	139	0.67	0.09	0.66	0.83
Lang. Model (LM)	42	0.43	**0.05**	0.55	0.76
Qual. Link (QL)	12	0.87	0.27	0.67	0.83
C ∪ L (baseline)	237	0.84	0.14	0.75	0.85
C ∪ L ∪ LM	279	0.87	0.11	0.81	0.86
C ∪ L ∪ QL	249	**0.92**	0.14	0.83	0.86
C ∪ L ∪ LM ∪ QL	291	0.89	0.10	**0.86**	**0.88**

TABLE IV
FEATURES, TRUE POSITIVE (TP) RATE, FALSE POSITIVE (TP) RATE, F-MEASURE (F), AND AREA UNDER ROC CURVE (AUC) FOR WEB SPAM CLASSIFIERS USING DIFFERENT FEATURE SETS ON DATASET. THE BEST SCORES ARE MARKED IN BOLD

WEBSPAM-UK2007					
Feature Set	Features	*TP*	*FP*	*F*	*AUC*
Content (C)	98	0.33	0.04	0.30	0.72
Link (L)	139	0.39	0.12	0.20	0.68
Lang. Model (LM)	42	0.24	**0.03**	0.24	0.72
Qual. Link (QL)	12	0.40	0.06	0.32	0.72
C ∪ L (baseline)	237	0.31	0.05	0.31	0.73
C ∪ L ∪ LM	279	0.33	**0.03**	0.33	0.75
C ∪ L ∪ QL	249	0.48	0.06	0.38	0.75
C ∪ L ∪ LM ∪ QL	291	**0.50**	0.06	**0.40**	**0.76**

establishing different costs to misclassifying. Before a model is learned on the training data, the data is reweighted to increase the sensitivity to spam cases.

Thus, as in Castillo *et al.* [6], we have imposed a zero cost to right predictions, and we have set to spam pages misclassified as nonspam a cost R times higher than nonspam pages misclassified as spam. Furthermore, as the aim of this work is to maximize the F-measure, we have looked for the value of R which maximize these measures. Fig. 8 illustrates the evolution of F-measure obtained by applying different costs to R. According to these results, we have set $R = 4$ in WEBSPAM-UK2006 dataset and $R = 14$ in WEBSPAM-UK2007.

C. Results

In order to check if the proposed features improve the precision of spam detection, we decided to use precomputed features available for the public dataset.[4] Specifically, we have used the content-based features and the transformed link-based features. In addition, we have combined different feature sets in order to obtain a classifier which has been able to detect both content-spam and link-spam cases. Finally, we have combined content, link, LM, and QL features, achieving a more accurate classifier. As a baseline for our experiments, we selected the precomputed content and link features in a combined way to detect different types of Web spam pages. These features were previously presented in [2] and [19].

The results of our experiments for WEBSPAM-UK2006 and WEBSPAM-UK2007 datasets are shown in Tables III and IV, respectively. As it can be seen, if we only use the precomputed features from datasets, we obtain the best results combining content and link-based features (C ∪ L). For this reason, we have chosen the union of these two sets of features as a baseline for our experiments.

Table III illustrates for the WEBSPAM-UK2006 dataset that QL features get an F-measure higher (0.67) than content (0.63), links (0.66), or LM features (0.55). This result is remarkable since the number of features used by this approach is much smaller (12) than content (98) or link-based features (139). Even so, QL features are not as efficient for themselves as the combination of content and link features (0.75). On the other hand, when we combine the baseline with the LM and QL-based features, we get several significant improvements. Specifically, the classifier using the combination of baseline and LM (C ∪ L ∪ LM) gets an improvement of 6% in the F-measure. Moreover, the combination of baseline and QL (C ∪ L ∪ QL) improves 8%, from 0.75 to 0.83. The most important observation is that if we consider the baseline, the classifier improves 11% in the F-measure, from 0.75 to 0.86, by combining the baseline, LM, and QL features (C ∪ L ∪ LM ∪ QL).

The detection rate is lower for the WEBSPAM-UK2007 dataset than in the previous dataset, as shown in Table IV. In this case, the collection is not well-balanced, having 4012 hosts labeled as "normal" and 154 as "spam." For this reason, the detection of spam is more difficult now and results are far worse than the WEBSPAM-UK2006 dataset. In spite of this problem, experiments show consistent results compared to the improvements obtained in the previous dataset. In any case, Table IV illustrates that QL features get an F-measure higher (0.32) than content (0.30), links (0.20), or LM features (0.24).

[4] Available: http://webspam.lip6.fr

As in the previous dataset, when we combine the baseline with the LM and QL-based features, we get several significant improvements. Specifically, the classifier using the combination of baseline and LM $(\mathrm{C} \cup \mathrm{L} \cup \mathrm{LM})$ gets an improvement of 2% in the F-measure. Moreover, the combination of baseline and QL $(\mathrm{C} \cup \mathrm{L} \cup \mathrm{QL})$ improves 7%, from 0.31 to 0.38. On the other hand, the main achievement is the classifier improves 9% in the F-measure of the baseline, from 0.31 to 0.40, by combining the baseline, LM, and QL features $(\mathrm{C} \cup \mathrm{L} \cup \mathrm{LM} \cup \mathrm{QL})$.

We can conclude from the values shown in Tables III and IV, that noteworthy improvements are obtained by combining LM and QL features. The four sets of features produce the best results because each set focuses on a different type of spam and they have complementary characteristics. Thus, this combination manages to detect content spam, link spam, nepotistic links, and QLs. Moreover, if we consider the sets separately, each one of them has a different impact on the F-measure parameters. While QL gets the best Precision, it also gets the worst Recall. LM gets the worst Precision, but it gets the best Recall. Finally, the combination of the four sets gets a very high Precision, without affecting the Recall.

VI. Conclusion

In this paper, we proposed a new methodology to detect spam in the Web, based on an analysis of QLs and a study of the divergence between linked pages.

To use QLs and the LM features effectively, we proposed a robust classifier based on a cost-sensitive algorithm. We have evaluated our methodology using the public WEBSPAM-UK2006 and WEBSPAM-UK2007 datasets and we focus on the F-measure, using the proposed features in a separate and also in a combined way. It has been proven that QL features have obtained better results than precomputed content and link-based features, even with many fewer features. In addition, when we combine the four sets of features and we apply them to WEBSPAM-UK2006 and WEBSPAM-UK2007 datasets, the system detects 89.4% and 54.2% of the spam domains, with an F-measure of 0.86 and 0.40, respectively. Thus, an improvement of the F-measure of 11% and 9%, respectively, is obtained.

Therefore, the comparisons with precomputed features show that the proposed methodology yields much better performance, indicating that LMs and QLs can be used to detect Web spam effectively.

In future works, we would like to analyze the relationship between a page and those that point to it, and to measure the disagreement between new sources of information in order to improve the performance of the LM approach. The current system is not intended to be a real-time application, nevertheless one of our future works is to look for a way of reducing the execution time with a low impact on the performance. In particular, we will study the effect of reducing the number of pages retrieved for each link, and the amount of links analyzed per page.

References

[1] J. Abernethy, O. Chapelle, and C. Castillo, "Webspam identification through content and hyperlinks," in *Proc. Fourth Int. Workshop on Adversarial Information Retrieval on the Web (AIRWeb)*, Beijing, China, 2008, pp. 41–44.

[2] L. Becchetti, C. Castillo, D. Donato, S. Leonardi, and R. Baeza-Yates, "Link-based characterization and detection of web spam," in *Proc. 2nd Int. Workshop on Adversarial Information Retrieval on the Web (AIRWeb'06)*, Seattle, WA, 2006, pp. 1–8.

[3] A. A. Benczúr, I. Bíró, K. Csalogány, and M. Uher, "Detecting nepotistic links by language model disagreement," in *Proc. 15th Int. Conf. World Wide Web (WWW'06)*, New York, 2006, pp. 939–940, ACM.

[4] A. A. Benczúr, K. Csalogány, T. Sarlós, and M. Uher, "Spamrank—Fully automatic link spam detection," in *Proc. First Int. Workshop on Adversarial Information Retrieval on the Web (AIRWeb*, Chiba, Japan, 2005, pp. 25–38.

[5] K. Bharat and M. R. Henzinger, "Improved algorithms for topic distillation in a hyperlinked environment," in *Proc. 21st Annu. Int. ACM SIGIR Conf. Research and Development in Information Retrieval*, New York, 1998, pp. 104–111, ACM.

[6] C. Castillo, D. Donato, L. Becchetti, P. Boldi, S. Leonardi, M. Santini, and S. Vigna, "A reference collection for web spam," *SIGIR Forum*, vol. 40, no. 2, pp. 11–24, 2006.

[7] C. Castillo, D. Donato, A. Gionis, V. Murdock, and F. Silvestri, "Know your neighbors: Web spam detection using the web topology," in *Proc. 30th Annu. Int. ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR'07)*, New York, 2007, pp. 423–430, ACM.

[8] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley-Interscience, 1991.

[9] N. Craswell, D. Hawking, and S. Robertson, "Effective site finding using link anchor information," in *Proc. 24th Annu. Int. ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR'01)*, New York, 2001, pp. 250–257, ACM.

[10] B. Davison, Recognizing Nepotistic Links on the Web 2000 [Online]. Available: citeseer.ist.psu.edu/davison00recognizing.html

[11] N. Eiron and K. S. McCurley, "Analysis of anchor text for web search," in *Proc. 26th Annu. Int. ACM SIGIR Conf. Research and Development in Informaion Retrieval (SIGIR'03)*, New York, 2003, pp. 459–460, ACM.

[12] O. Gospodnetic and E. Hatcher, *Lucene in Action*. Greenwich, CT: Manning, 2004.

[13] Z. Gyöngyi and H. Garcia-Molina, "Web spam taxonomy," in *Proc. First Int. Workshop on Adversarial Information Retrieval on the Web (AIRWeb)*, 2005 [Online]. Available: citeseer.ist.psu.edu/gyongyi05web.html

[14] R. Jin, A. G. Hauptmann, and C. X. Zhai, "Title language model for information retrieval," in *Proc. 25th Annu. Int. ACM SIGIR Conf. Research and Development in Information Retrieval*, New York, 2002, pp. 42–48, ACM.

[15] J. Martinez-Romo and L. Araujo, "Recommendation system for automatic recovery of broken web links," in *IBERAMIA*, 2008, pp. 302–311.

[16] J. Martinez-Romo and L. Araujo, "Retrieving broken web links using an approach based on contextual information," in *Proc. 20th ACM Conf. Hypertext and Hypermedia (HT'09)*, New York, 2009, pp. 351–352, ACM.

[17] J. Martinez-Romo and L. Araujo, "Web spam identification through language model analysis," in *Proc. Fifth Int. Workshop on Adversarial Information Retrieval on the Web (AIRWeb)*, Madrid, Spain, 2009, pp. 21–28, ACM.

[18] G. Mishne, D. Carmel, and R. Lempel, "Blocking blog spam with language model disagreement," in *Proc. First Int. Workshop on Adversarial Information Retrieval on the Web (AIRWeb)*, Chiba, Japan, 2005, pp. 1–6.

[19] A. Ntoulas, M. Najork, M. Manasse, and D. Fetterly, "Detecting spam web pages through content analysis," in *Proc. 15th Int. Conf. World Wide Web (WWW'06)*, New York, 2006, pp. 83–92, ACM.

[20] J. M. Ponte and W. B. Croft, "A language modeling approach to information retrieval," in *Proc. 21st Annu. Int. ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR'98)*, New York, 1998, pp. 275–281, ACM.

[21] X. Qi, L. Nie, and B. D. Davison, "Measuring similarity to detect qualified links," in *Proc. 3rd Int. Workshop on Adversarial Information Retrieval on the Web (AIRWeb'07)*, New York, 2007, pp. 49–56, ACM.

[22] S. Webb, J. Caverlee, and C. Pu, "Predicting web spam with http session information," in *Proc. 17th ACM Conf. Information and Knowledge Management (CIKM'08)*, New York, 2008, pp. 339–348, ACM.

[23] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed. San Mateo, CA: Morgan Kaufmann, 2005 [Online]. Available: /bib/private/witten/Data Mining Practical Machine Learning Tools and Techniques 2d ed—Morgan Kaufmann.pdf

Lourdes Araujo received the B.S. degree in physics and the Ph.D. degree in computer science from the Universidad Complutense de Madrid, Spain, in 1987 and 1994, respectively.

She is currently Computer Science Professor at the UNED in Madrid, Spain, and belongs to the Natural Language Processing and Information Retrieval Group of this University. Her current research interests include natural language processing and information retrieval as well as evolutionary algorithms.

Juan Martinez-Romo received the B.S. and M.S. degrees from the Universidad Rey Juan Carlos at Madrid (URJC), Spain, in 2003 and 2005, respectively. He is currently working toward the Ph.D. degree at UNED in Madrid, Spain.

From 2005 to 2007, he developed his research activity at the Grupo de Sistemas y Comunicaciones at URJC. He is currently Assistant Professor at the UNED in Madrid, Spain, and belongs to the NLP & IR Group of this University. His research interests include web information retrieval, natural language processing, and web spam.