Analyzing Information Retrieval Methods to Recover Broken Web Links*

Juan Martinez-Romo and Lourdes Araujo

NLP & IR Group, UNED, Madrid 28040, Spain
{juaner,lurdes}@lsi.uned.es

Abstract. In this work we compare different techniques to automatically find candidate web pages to substitute broken links. We extract information from the anchor text, the content of the page containing the link, and the cache page in some digital library. The selected information is processed and submitted to a search engine. We have compared different information retrieval methods for both, the selection of terms used to construct the queries submitted to the search engine, and the ranking of the candidate pages that it provides, in order to help the user to find the best replacement. In particular, we have used term frequencies, and a language model approach for the selection of terms; and cooccurrence measures and a language model approach for ranking the final results. To test the different methods, we have also defined a methodology which does not require the user judgments, what increases the objectivity of the results.

Keywords: Information retrieval, link integrity, recommender system.

1 Introduction

Missing pages are very frequent on the web: many websites disappear while others are not properly maintained. Thus, broken links represent an important problem that affects the information access and the ranking of the search engines. There exist several validators to check the broken links of our pages. However, once we have detected a broken link, it is not always easy nor fast to find again the disappeared page. In this work, we try to analyze the recovery of broken links as an information retrieval problem for building an user recommendation system. In the same way that an user query to a search engine to find a Web page with the information he needs, our system captures all data related to the broken link context in order to find a new page that containing information similar to the missing one.

Our system checks the links of the page given as input. For those which are broken, the system proposes to the user a set of candidate pages to substitute the broken link. Figure 1 presents a scheme of the proposed system. The first step of our work has been the analysis of a large number of web pages and their

* This work has been partially supported by the Spanish Ministry of Science and Innovation within the project QEAVis-Catiex (TIN2007-67581-C02-01) and the Regional Government of Madrid under the Research Network MAVIR (S-0505/TIC-0267).

C. Gurrin et al. (Eds.): ECIR 2010, LNCS 5993, pp. 26–37, 2010.

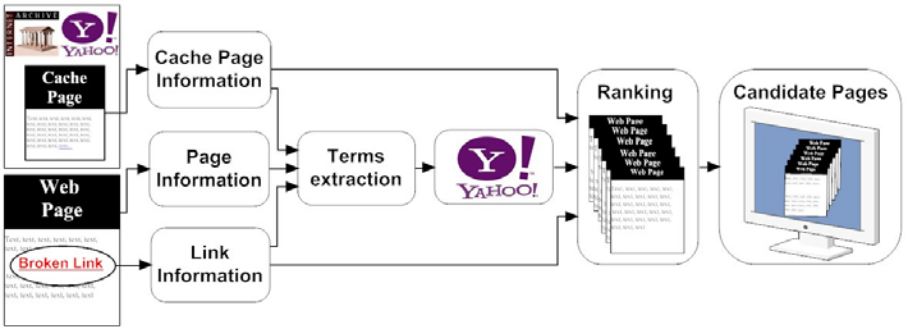

Fig. 1. Scheme of the system for automatic recovering of broken links

links in order to determine which ones are the most useful sources of information and which of them are the most appropriate in each case. Sometimes we can recover a broken link by entering the anchor text as an user query in a search engine. There are many works which have analyzed the importance of the anchor text and title like a source of information[1]. However, there are many cases in which the anchor text does not contain enough information to do that. In these cases, we can compose queries adding terms extracted from other sources of information to the anchor text of the broken link. For that, our system performs a form of query expansion[2], a well-known method to improve the performance of information retrieval systems. In our case, the original query is composed of the terms extracted from the anchor text, and the sources of expansion terms are the elements of the parent web page containing the broken link (text, title, url, etc), and also, if they exist, the elements of the cache page corresponding to the disappeared page that can be stored in a search engine (*Yahoo*) or web archive (*Wayback Machine*). In this work we have investigated the performance of different approaches to extract the expansion terms from the mentioned sources of information in the context of a link.

After the term extraction step, different expanded queries are submitted to the considered search engine, and the set of top ranked documents are retrieved. In order to tune the results, the pages recovered in this way are ranked according to relevance measures obtained by applying information retrieval (IR) techniques, and finally they are presented to the user.

In order to evaluate the different IR techniques considered, we have developed a methodology which mainly relies on the random selection of pages and the use of links that are not really broken to check how many are properly recovered.

The rest of this paper is organized as follows. We start off reviewing related work in the next section 2; section 3 describes the methodology we have followed; section 4 analyses how useful is the anchor text of the links to recover the page; section 5 studies the suitability of different sources of information to provide terms for the query expansion process; section 6 is devoted to describe the process to rank the candidate documents; section 7 presents the scheme resulting of the previous analysis; section 8 analyses the parameter setting of the proposed algorithm. Finally, section 9 draws the main conclusions of this work.

2 Background and Related Work

Despite the problem of broken links was considered the second most serious problem on the Web[3] many years ago, missing pages are still frequent when users are surfing the Internet. Previous works quantified this problem: Kahle[4] reported the expected life-time of a web page is 44 days. Koehler et al.[5] performed a longitudinal study of web page availability and found the random test collection of Urls eventually reached a "steady state" after approximately 67% of the Urls were lost over a 4-year period. Markwell et al.[3] monitored the resources of three authentic courses during 14 months, and 16.5% of the links had disappeared or were nonviable.

Most of previous attempts to recover broken links are based on information annotated in advance with the link[6,7]. Closest to our research[8], Nakamizo et al.[9,10] have developed a tool that finds new Urls of web pages after pages are moved. This tool outputs a list of web pages sorted by their plausibility of being link authorities. Klein et al.[11] save a small set of terms (lexical signature) derived from a document that capture the "aboutness" of that document, and they can be used to discover a similar Web page if it disappears in the future. Harrison et al.[12] presented Opal, a framework for interactively locating missing web pages, using the previous lexical signatures which are then used to search for similar versions of a Web page. Our work differs from previous proposals since it does not rely on any information about the links annotated in advance, and it can be applied to any web page. Previous works do not present a evaluation methodology, thus we have also defined a methodology which does not require the user judgments, what increases the objectivity of the results.

Many works appeared using different techniques in the TREC-10 Web Track (homepage finding task) to perform the proposed task, although our work differs in the following respects: (i) the size of the collection is much smaller than the whole Internet, (ii) the number of potential candidates is much more reduced, and (iii) most of papers used the Url depth as a main factor for the page selection.

3 Methodology

If we analyze the usefulness of the different sources of information directly employed on broken links, it is very difficult to evaluate the quality of the sources of information and the techniques used to extract the terms. Therefore, at this phase of analysis, we employ random web links, which are not really broken, and we called *pseudobroken* links. Thus we have the page at which they point and we are able to evaluate our recommendation. We take links from pages selected randomly (250 words are required) by means of successive requests to *www.randomwebsite.com*, a site that provides random web pages. After the best sources of information and information retrieval techniques are selected, we employ really broken links in order to test the whole system.

We consider that a link has been recovered if the system finds a web page which content is practically the same that the *pseudobroken* link. For that, we apply the vector space model [13], i.e we represent each page by a term vector

and calculate the cosine distance between them (similarity). If this value is higher than 0.9, we consider that the page has been recovered. Lowering the similarity threshold (e.g. 0.8) very few additional links are recovered, and the number of wrong results increase.

4 Analyzing the Anchor Text in a Hyperlink

In many cases, terms which compose the anchor text of a hyperlink are the main source of information to identify the pointed page. To verify this theory we performed a study searching in Yahoo for the anchor text. Using the previously defined similarity, 41% of the links were recovered in the top ten results. In addition, 66% of the recovered links appear in the first position. These results prove that anchor text is a efficient source of information to recover a broken link.

Sometimes anchor terms provide little or no descriptive value. Let us imagine a link whose anchor text is "click here". In this case, finding the broken link might be impossible. For this reason it is very important to analyze these terms so as to be able to decide which tasks should be performed depending on their quantity and quality. Thus, the system carry out a recognition of named entities (persons, organizations or places) on the anchor text in order to extract certain terms whose importance is higher than the remaining ones. Several experiments have proved that the presence of any named entity in the anchor favors the recovery of the link. The most prominent result is the very small number of cases in which the correct document is recovered when the anchor consists of just a term and it is not a named entity.

5 The Page Text

The most frequent terms of a web page (after removing stop-words) are a way to characterize the main topic of the cited page. This technique requires the page text to be long enough. A clear example of utility of this information are the links to personal pages. The anchor of a link to a personal page is frequently formed by the name of that person. However, in many cases forename and surname do not identify a person in a unique way, specially if they are very common. If we perform a search using only the forename and the surname, the personal page of this person probably will not appear among the first pages retrieved. However, if we expand the query using some terms related to that person, that can be present at his web page, then his personal web page will go up to the first positions.

We have applied classical information retrieval techniques, described below, to extract the most representative terms from a page. After eliminating the stop words, we generate a ranked term list. The first ten terms of this list are used to compose ten expanded queries, one for each expansion term. Every query is formed by the anchor text and it is expanded with each of those terms. Finally, the first ten retrieved documents are taken in each case.

5.1 Frequency-Based Approaches to Select Terms

Frequency-Based are the most simple approaches to select expansion terms. We have considered two different criteria based on frequencies for term selection. The first one is the raw term frequency (TF) in the parent or cache page. There are some terms with very little or no discriminating power as descriptors of the page, despite they are frequent on it. The reason is that those terms are also frequent in many other documents of the considered collection. To take into account these cases, we apply the well-known *Tf-Idf* weighting scheme for a term, where *Idf*(t) is the inverse document frequency of that term. A dump of English Wikipedia articles[1] has been used as reference collection.

5.2 Language Model Approach

One of the most successful methods based on term distribution analysis uses the concept of Kullback-Liebler Divergence[14] to compute the divergence between the probability distributions of terms in the whole collection and the particular considered documents. The most suitable terms to expand the query are those with a high probability in the document, which is the source of terms, and a low probability in the whole collection. For the term t this divergence is:

$$KLD_{(PP,PC)}(t) = P_P(t)\log\frac{P_P(t)}{P_C(t)} \tag{1}$$

where $P_P(t)$ is the probability of the term t in the considered page, and $P_C(t)$ is the probability of the term t in the whole collection.

Computing this measure requires a reference collection of documents. The relation between this reference collection and the analyzed document, is an important factor in the results obtained with this approach. Obviously, we can not use the whole web as reference collection. To study the impact of this factor on the results we have used two different collections of web pages indexed with Lucene: (i) English Wikipedia articles (3.6 million) dump (Enwiki) and (ii) Homepage at sites (4.5 million) in DMOZ Open Directory Project (ODP).

5.3 Comparing Different Approaches for the Extraction of Terms

Figure 2 show the results obtained using frequency, *Tf-Idf* and *KLD* for the extraction of expansion terms from the parent page and the cache page respectively. In the case of the parent page (Figure 2(i)) the frequency method performs slightly better than *Tf-Idf*, whereas for the cache page (Figure 2(ii)) it is the opposite. We can observe that when we use *KLD*, the results obtained with the Wikipedia as reference collection are better (the total number of correct recovered pages). The reason is probably that this collection provides a wider range of topics. Another important observation is that for extracting terms from the parent page, the results obtained with the methods based on frequencies are higher than *KLD*. On

[1] http://download.wikimedia.org/enwiki/

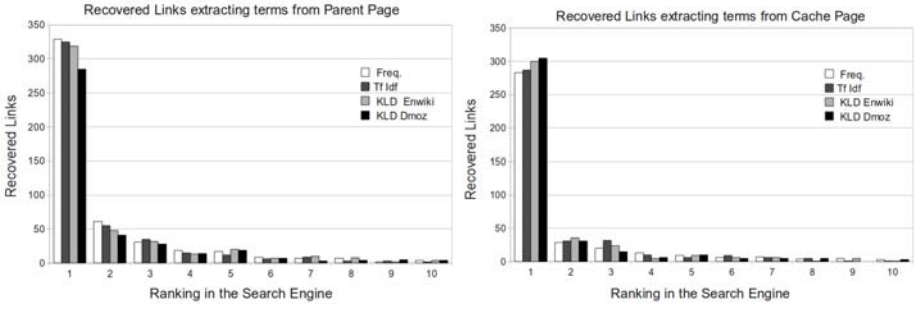

Fig. 2. Recovered links by expanding the query with (i) terms from the parent page and (ii) terms from the cache page. The data of the figure indicate the number of cases in which Yahoo has provided the correct page in the corresponding position when the system expands with the terms extracted applying different approaches.

Table 1. Analysis of the number of retrieved documents in the first position and in top 10 positions, according to if query expansion is used or not

Analysis	**1st position**	**1-10 positions**
No Expansion (anchor text)	253	380
Expansion (terms from page)	213	418

the contrary, *KLD* is the most suitable approach to extract terms from the cache. The reason is perhaps that in some cases the parent page content is not closely related to the page to recover, and thus, by refining the methods to select the most representative terms of the parent page does not improve the results. Accordingly, we have used the frequency method for extracting terms from the parent page, and we have used *KLD* with the English Wikipedia as reference collection for extracting terms from the cache page in the remaining experiments.

5.4 The Effect of Expansion on the Relationship between Precision-Recall

We performed a new experiment in order to study the effect of expansion in the relationship between precision-recall. In Table 1 can be observed that expansion considerably increases the number of recovered links ranked in the top ten positions (recall). In spite of this, the number of recovered links ranked in the first position is reduced (precision). Accordingly, we think that the most suitable mechanism is to apply both recovery approaches, and later ranking the whole set of results to present the user the most relevant web pages ranked in the top positions.

6 Ranking Methods

Once the system has retrieved a set of candidate pages to replace the broken link by combining both *No Expansion* and *Expansion* approaches, the system

needs to present the results to the user in decreasing order of relevance. In order to establish the best ranking function for the candidate pages, we performed an analysis in order to compare different similarity approaches and elements from parent, cache and candidate pages.

6.1 Vector Space Model

The vector space model[13] is one of the approaches that we have applied to represent the documents. Methods based on term cooccurrence[15] have been used very frequently to identify semantic relationships among documents. In our experiments, we have used the well-known Tanimoto, Dice and Cosine cooccurrence coefficients to measure the similarity between the vectors representing the reference document D_1 and the candidate document D_2:

$$\text{Tanimoto}(\boldsymbol{D}_1, \boldsymbol{D}_2) = \frac{\boldsymbol{D}_1\boldsymbol{D}_2}{|\boldsymbol{D}_1|^2 + |\boldsymbol{D}_2|^2 - \boldsymbol{D}_1\boldsymbol{D}_2} \tag{2}$$

$$\text{Dice}(\boldsymbol{D}_1, \boldsymbol{D}_2) = \frac{2\boldsymbol{D}_1\boldsymbol{D}_2}{|\boldsymbol{D}_1|^2 + |\boldsymbol{D}_2|^2} \tag{3}$$

$$\text{Cosine}(\boldsymbol{D}_1, \boldsymbol{D}_2) = \frac{\boldsymbol{D}_1\boldsymbol{D}_2}{|\boldsymbol{D}_1||\boldsymbol{D}_2|} \tag{4}$$

6.2 Language Model Approach

We have also applied a language model approach to rank the set of candidate documents. In this case we look at the differences in the term distribution between two documents by computing the Kullback-Leibler divergence:

$$KLD(D_1||D_2) = \sum_{t \in D_1} P_{D_1}(t) log \frac{P_{D_1}(t)}{P_{D_2}(t)} \tag{5}$$

where $P_{D_1}(t)$ is the probability of the term t in the reference document, and $P_{D_2}(t)$ is the probability of the term t in the candidate document.

6.3 Performance of Ranking Methods

We have applied the approaches described above to different elements from the parent, cache and candidate pages. Specifically, we have studied the similarity among the following pairs of elements: (i) parent anchor text & candidate title, (ii) parent anchor text & candidate content, (iii) parent content & candidate content, and (iv) cache content & candidate content.

In addition to these comparisons, we also used the anchor text and the snippet of the candidate document but the results were not improved. We can observe in Figure 3(i) and Figure 3(ii) that the results obtained with *KLD* are worse than those obtained by using the cooccurrence measures, especially in Figure 3(i). In these Figures we are studying the similarity between a very short text, the anchor

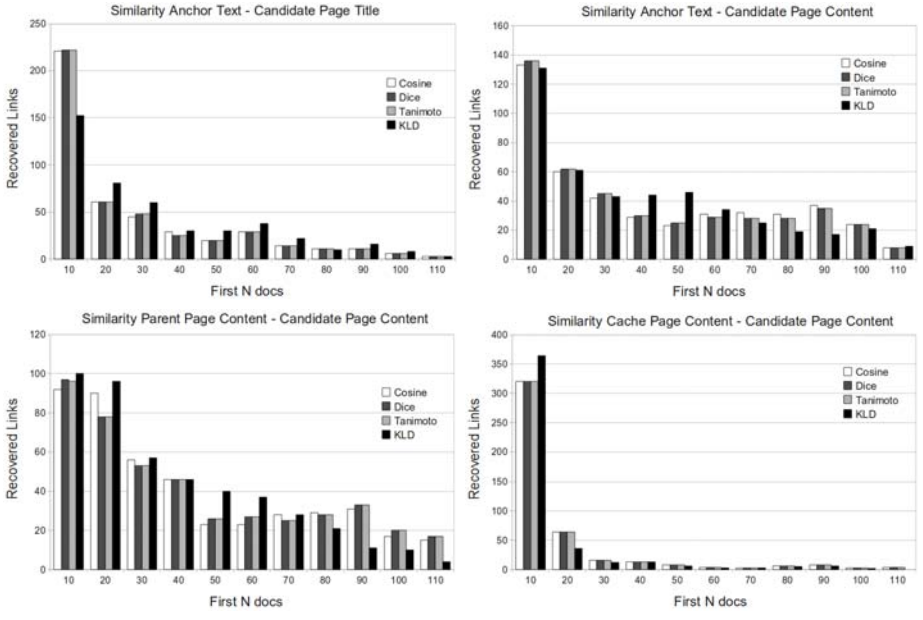

Fig. 3. Results of different approaches (Cosine, Dice, Tanimoto and Kullback-Liebler Divergence) applied to measure the similarity between: (i) the Anchor Text of the broken link and the Title of the candidate page, (ii) the Anchor Text of the broken link and the Content of the candidate page, (iii) the Content of the page where is the broken link and the Content of the candidate page, and (iv) the Content of the Cache page of the broken link and the Content of the candidate page. Results show the position of the best page recovered by the system after the ranking.

text, and other short text which is the title of the candidate page (Figure 3(i)) or the parent page content (Figure 3(ii)). On the contrary, *KLD* performs better than the cooccurrence measures in Figure 3(iii) and Figure 3(iv), where we are measuring the similarity between the content of two pages; the parent or the cache page, and the candidate page. Thus, we can conclude that *KLD* performs better that the cooccurrence methods only if it is applied to texts long enough, such as the content of a page. In Figure 3(iv) we can observe that, as expected, the best results are obtained when they are ranked by similarity with the cache page. However, in many cases this page is not available (near a 60%). Regarding the remaining results, it can be observed that the best results are obtained when the similarity between the anchor text of the broken link and the title of the candidate page is used, and by applying cooccurrence methods. According to these results, *KLD* will be used to rank the candidates pages with respect to the cache page if the system is able to retrieve the cache page. Otherwise, the system will use the similarity between the anchor text and the candidate page title, measured with a cooccurrence method, such a *Dice*, which performs slightly better in some cases.

7 Algorithm for Automatic Recovery of Links

The results of the analysis described in the previous sections suggest several criteria to decide for which cases there is enough information to try the retrieval of the link and which sources of information to use. According to them, we propose the recovery process which appears in Figure 4. First of all, it is checked whether the anchor number of terms is just one (`length(anchor) = 1`) and whether it does not contain named entities (`NoNE(anchor)`). If both features are found, the retrieval is only attempted provided the link of the missing page appears in the cache of a search engine or web archive (`InCache(page)`), and therefore we have reliable information to verify that the proposal presented to the user can be useful. Otherwise, the user is informed that the recommendation is not possible (`No_recovered`). If the page is in the cache, then the recovery is performed, expanding the query (anchor terms) with terms extracted from the cache using *KLD*. Then the results are ranked (by similarity between the candidate page and the cache page computed with *KLD*) and only if any of them is sufficiently similar to the cache content (`similarity(docs, cache(page) > 0.9)`), the user is recommended this list of candidate documents. In the remaining cases, that is, when the anchor has more than one term or when it contains some named entity, the recovery is performed using the anchor terms and the terms from the cache (applying *KLD*) or parent page (applying frequency selection). After that, all documents are grouped and ranked according to the cache page (`rank(docs, cache_content_KLD)`) if it is available in a search engine or web archive, or according to the similarity between the anchor text and the title of the candidate page applying the Dice cooccurrence coefficient (`rank(docs, anchor_title_Dice)`) otherwise.

7.1 Results Applying the System to Broken Links

The previous algorithm has been applied to a set of pages with broken links, but they have only been used those that were present in a digital library (search engine cache or web archive). The reason is that only in this case we can objectively evaluate the results. Thanks to the algorithm, the system recovered 553 from 748 broken links (74% of the total links). Table 2 shows a ranking of recovered links. We have verified that in some cases the original page is found (it has been moved to other web site) and in some other cases, the system retrieved pages with very similar content. We can observe the system is able to provide useful replacements documents among the top 10 positions in 46% of the recovered broken links, and among the 20 first ones in 70% of the cases.

8 Tuning the Algorithm Parameters

An important issue to investigate is the trade-off between the amount of information collected to recover the broken links, and the time required to do it. We can expect that, in general, the more information available, the better the

```
if length(anchor) = 1 and NoNE(anchor) then
   if InCache(page) then
     docs = web_search(anchor + cache_terms_KLD)
     rank(docs, cache_content_KLD)
     if similarity(docs, cache(page) > 0.9) then
         user_recommendation(docs)
     else
         No_recovered
   else
     No_recovered
else
   docs = web_search(anchor)
   if InCache(page) then
     docs = docs + web_search(anchor + cache_terms_KLD)
     rank(docs, cache_content_KLD)
   else
     docs = docs + web_search(anchor + page_terms_Freq)
     rank(docs, anchor_title_Dice)
   user_recommendation(docs)
```

Fig. 4. Links Automatic Recovery Algorithm for broken links

Table 2. Number of recovered broken links (best candidate which content is very similar to the missing page) according to his cache similarity, among first N documents using the proposed algorithm

First N documents	Recovered Broken Links
1-10	254
10-20	134
20-50	122
50-110	43

recommendation for the broken link. However it is important to know the cost incurred for each increment of the collected data. The amount of information collected mainly depends on two parameters: the number of terms used to expand the query extracted from the anchor text, and the number of hits taken from the results of the search engine for each query. We performed several experiments to evaluate how these parameters affect the results. Figure 5(a) shows the number of recovered links according to the number of retrieved hits from the search engine. We can observe that the improvement is much smaller from 25-30 hits, especially when the expansion is not performed. Figure 5(b) shows the number of recovered links for different number of terms used in the query expansion process. In these experiments the number of hits retrieved from the search engine was set to 10. It is interesting to notice that the expansion approach beats the approach without expansion when expanding with 6 or more terms, and this improvement is quite small from 10 or 15 terms.

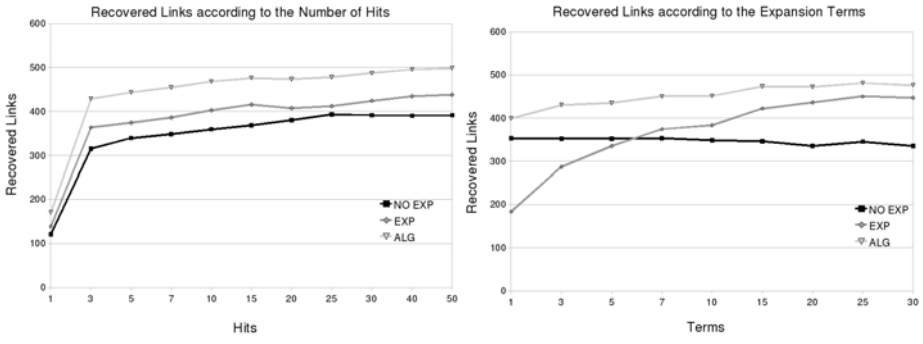

Fig. 5. Recovered Links according to the number of hits/terms used to carry out each query/expanded query. This figure shows the results using a method without expansion (NO EXP), an expansion method (EXP) and the combination of both methods (ALG).

Finally we have carried out an execution time analysis in order to determine the influence on time of the number of hits retrieved from the search engine and the number of terms used for expansion. According to obtained results and the previous results showed in Figure 5, the number of terms and hits has been set to 10, as a trade-off between the improvement in the performance and the execution time (around 17 seconds for every link).

9 Conclusions

In this paper we have analyzed different information retrieval methods for both, the selection of terms used to construct the queries submitted to the search engine, and the ranking of the candidate pages that it provides, in order to help the user to find the best replacement for a broken link. To test the sources, we have also defined a evaluation methodology which does not require the user judgments, what increases the objectivity of the results. We have also studied the effect of using terms from the page that contains the link, and a cache page stored in some search engine or web archive to expand the query formed by the anchor text. This study shows that the results are better when the query is expanded, than when the anchor text is used alone. Thus, query expansion reduce the ambiguity that would entail the limited quantity of anchor terms. We have compared different methods for the extraction of terms to expand the anchor text. Experiments have shown that the best results are obtained using a frequency approach if the cache page is not available, and a language model approach otherwise. We have decided to combine both methods and later reordering the obtained results by applying a relevance ranking in order to present to the user the best candidate pages at the beginning. We have also compared different approaches to rank the candidate documents: cooccurrence approaches and a language model divergence approach (*KLD*). The best results are obtained applying *KLD* when the cache page is available, otherwise a cooccurrence method

such as Dice applied between the anchor text of the broken link and the title of the candidate page. This analysis has allowed us to design a strategy that has been able to recover a page very similar to the missing one in 74% of the cases. Moreover, the system was able to provide 46% from those recovered links in the top ten of the results, and among the top 20 in 70%.

References

1. Craswell, N., Hawking, D., Robertson, S.: Effective site finding using link anchor information. In: SIGIR 2001: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 250–257. ACM Press, New York (2001)
2. Efthimiadis, E.N.: Query expansion. Annual Review of Information Systems and Technology 31, 121–187 (1996)
3. Markwell, J., Brooks, D.W.: Broken links: The ephemeral nature of educational www hyperlinks. Journal of Science Education and Technology 11(2), 105–108 (2002)
4. Kahle, B.: Preserving the internet. Scientific American 276(3), 82–83 (1997)
5. Koehler, W.: Web page change and persistence—a four-year longitudinal study. J. Am. Soc. Inf. Sci. Technol. 53(2), 162–171 (2002)
6. Ingham, D., Caughey, S., Little, M.: Fixing the broken-link problem: the w3objects approach. Comput. Netw. ISDN Syst. 28(7-11), 1255–1268 (1996)
7. Shimada, T., Futakata, A.: Automatic link generation and repair mechanism for document management. In: HICSS 1998: Proceedings of the Thirty-First Annual Hawaii International Conference on System Sciences, vol. 2, p. 226. IEEE Computer Society Press, Los Alamitos (1998)
8. Martinez-Romo, J., Araujo, L.: Recommendation system for automatic recovery of broken web links. In: Geffner, H., Prada, R., Machado Alexandre, I., David, N. (eds.) IBERAMIA 2008. LNCS (LNAI), vol. 5290, pp. 302–311. Springer, Heidelberg (2008)
9. Nakamizo, A., Iida, T., Morishima, A., Sugimoto, S., Kitagawa, H.: A tool to compute reliable web links and its applications. In: SWOD 2005: Proc. International Special Workshop on Databases for Next Generation Researchers, pp. 146–149. IEEE Computer Society, Los Alamitos (2005)
10. Morishima, A., Nakamizo, A., Iida, T., Sugimoto, S., Kitagawa, H.: Pagechaser: A tool for the automatic correction of broken web links. In: ICDE, pp. 1486–1488 (2008)
11. Klein, M., Nelson, M.L.: Revisiting lexical signatures to (re-)discover web pages. In: Christensen-Dalsgaard, B., Castelli, D., Ammitzbøll Jurik, B., Lippincott, J. (eds.) ECDL 2008. LNCS, vol. 5173, pp. 371–382. Springer, Heidelberg (2008)
12. Harrison, T.L., Nelson, M.L.: Just-in-time recovery of missing web pages. In: HYPERTEXT 2006: Proceedings of the seventeenth conference on Hypertext and hypermedia, pp. 145–156. ACM Press, New York (2006)
13. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval. Cambridge University Press, Cambridge (2008)
14. Cover, T.M., Thomas, J.A.: Elements of information theory. Wiley Interscience, New York (1991)
15. Rijsbergen, C.J.V.: A theoretical basis for the use of cooccurrence data in information retrieval. Journal of Documentation 33, 106–119 (1977)