Retrieving Broken Web Links using an Approach based on Contextual Information *

Juan Martinez-Romo
NLP & IR Group at UNED
28040 Madrid, Spain
juaner@lsi.uned.es

Lourdes Araujo
NLP & IR Group at UNED
28040 Madrid, Spain
lurdes@lsi.uned.es

ABSTRACT

In this short note we present a recommendation system for automatic retrieval of broken Web links using an approach based on contextual information. We extract information from the context of a link such as the anchor text, the content of the page containing the link, and a combination of the cache page in some search engine and web archive, if it exists. Then the selected information is processed and submitted to a search engine. We propose an algorithm based on information retrieval techniques to select the most relevant information and to rank the candidate pages provided for the search engine, in order to help the user to find the best replacement. To test the different methods, we have also defined a methodology which does not require the user judgements, what increases the objectivity of the results.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing

General Terms

Design, Algorithms, Measurement

Keywords

information retrieval, link integrity, recommender system

1. INTRODUCTION

No matter how well built a website is, there is no avoiding the fact that sometimes the page a user is looking for just is missing. What matters is how this website resolves the problem when it occurs. Missing pages represent an important problem that affects the information access and the ranking of the search engines. Most of previous attempts to recover broken links are based on information annotated in advance with the link[5, 2, 4]. Thought with a purpose different to repairing broken links, other works[3] have investigate mechanisms to extract information from the links and the context they appear in. Some of these mechanisms have been tested in our system for recovering broken links. 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.

Our system (Figure 1) 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. The candidate pages are obtained by submitting to a search engine queries composed of terms extracted from different sources. In our case, the original query are 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, 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*). After the term extraction step, different 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. The resulting list of pages is 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.

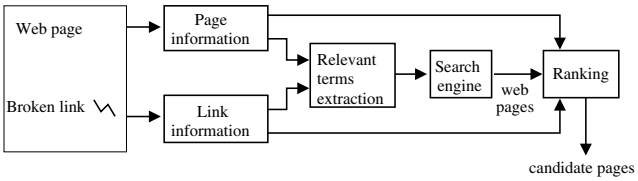

Figure 1: Scheme of the system for automatic recovering of broken links.

2. SOURCES OF INFORMATION

In many cases the words which compose the anchor text of a hyperlink are the main source of information to identify the pointed page. There are many works which have analyzed the importance of the Anchor Text like a source of information. Eiron and McCurley[1] carried out a study

*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).

HT'09, June 29–July 1, 2009, Torino, Italy.
ACM 978-1-60558-486-7/09/06.

where compared the usefulness of the anchor text and the content page in a web search task. McBryan[3] proposed to use the anchor text as a help to the search of web resources.

A combination of different tests to consider that a link has been recovered has been used. If the Urls do not match, we apply the vector space model, 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.

The most frequent terms of a web page are a way to characterize the main topic of the cited page. We have applied classical information retrieval techniques to extract the most representative terms from a page. Some of these approaches are based on term frequencies, while others are language modeling approaches based on the differences between the probability distribution of terms in a collection and in the considered parent or cache page. The first ten terms of this list are used to expand the query formed by the anchor text, i.e. the query is expanded with each of those terms, and the first ten documents retrieved in each case are taken.

3. RANKING CANDIDATE LINKS

At this point we have retrieved a set of candidate pages to replace the broken link. These pages are the results of searching in the web with the anchor and with the anchor plus terms from the parent page. Now we want to present the results to the user in decreasing order of relevance. To calculate this relevance we have considered two sources of information. First, if it exists, we use the cache page pointed to by the broken link stored in a search engine or web archive. If this information does not exist, then we use the parent page which contains the broken link. The idea is that the pointed page's topic will be related to one of the linked page. In both cases we have tested different ranking approaches, some of them based on the vector space model, and also a language model approach.

We have 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 computing the Kullback-Leibler divergence:

$$KLD(D_1||D_2) = \sum_{t \in D_1} P_{D_1}(t) log \frac{P_{D_1}(t)}{P_{D_2}(t)} \qquad (1)$$

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.

4. ALGORITHM

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 2.

We have applied this algorithm to a set of Web pages with broken links. Thanks to the algorithm, the system recovered 553 from 748 links (74% of the total links). We have verified that in some cases the original page is found (it has been moved to other Url). In some other cases, we have retrieved pages with very similar content. Moreover, the system is able to provide useful replacements documents between the first 10 positions in 46% of the recovered links, and between the 20 first ones in 70% of the cases.

```
if length(anchor) = 1 and NoNE(anchor) then
   if InCache(page) then
     docs = web_search(anchor + cache_terms)
     rank(docs, cache_content)
     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)
     rank(docs, cache_content)
   else
     docs = docs + web_search(anchor + page_terms)
     rank(docs, anchor_title)
   user_recommendation(docs)
```

Figure 2: Automatic Recovery Algorithm for broken links.

5. CONCLUSIONS

In this work we have analyzed different sources of information that we can use to carry out an automatic recovery of web links that are not valid anymore. Results indicate that the anchor terms can be very useful, especially if there are more than one and if they contain some named entity. We have also studied the effect of using terms, from the page that contains the link and the cache page pointed to by the broken link, for expanding the queries. In this way, we reduce the ambiguity that would entail the limited quantity of anchor terms.

The result of 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 is able to provide 46% from these recovered links in the top ten of the results, and between the top 20 in 70%.

6. REFERENCES

[1] N. Eiron and K. S. McCurley. Analysis of anchor text for web search. In *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 459–460, New York, NY, USA, 2003. ACM.

[2] T. L. Harrison and M. L. Nelson. Just-in-time recovery of missing web pages. In *HYPERTEXT '06: Proceedings of the seventeenth conference on Hypertext and hypermedia*, pages 145–156, New York, NY, USA, 2006. ACM.

[3] O. A. McBryan. GENVL and WWWW: Tools for Taming the Web. In O. Nierstarsz, editor, *Proceedings of the first International World Wide Web Conference*, page 15, CERN, Geneva, 1994.

[4] A. Morishima, A. Nakamizo, T. Iida, S. Sugimoto, and H. Kitagawa. Pagechaser: A tool for the automatic correction of broken web links. In *ICDE*, pages 1486–1488, 2008.

[5] A. Nakamizo, T. Iida, A. Morishima, S. Sugimoto, , and H. Kitagawa. A tool to compute reliable web links and its applications. In *SWOD '05: Proc. International Special Workshop on Databases for Next Generation Researchers*, pages 146–149. IEEE Computer Society, 2005.