Recommendation System for Automatic Recovery of Broken Web Links*

Juan Martinez-Romo and Lourdes Araujo

Dpto. Lenguajes y Sistemas Informáticos. UNED, Madrid 28040, Spain
juaner@lsi.uned.es, lurdes@lsi.uned.es

Abstract. In the web pages accessed when navigating throughout Internet, or even in our own web pages, we sometimes find links which are not valid any more. The search of the right web pages which correspond to those links is often hard. In this work we have analyzed different sources of information to automatically recover broken web links so that the user can be offered a list of possible pages to substitute that link. Specifically, we have used either the anchor text or the web page containing the link, or a combination of both. We report the analysis of a number of issues arising when trying to recover a set of links randomly chosen. This analysis has allowed us to decide the cases in which the system can perform the retrieval of some pages to substitute the broken link. Results have shown that the system is able to do reliable recommendations in many cases, specially under certain conditions on the anchor text and the parent page.

Keywords: information retrieval, World Wide Web, broken links.

1 Introduction

The web is a highly dynamic system with a continuous creation, deletion and movement of web pages. This often causes page links to become broken some time after the page creation. We now and then find this situation in Internet. This problem also forces us to verify frequently our own pages to check if their links are still valid. The search of the new location of a page that has been moved, or of a new page whose contain is similar to a disappeared page, is sometimes a difficult task. In the case of our own pages, the task can be easier, but still tedious.

There have been some attempts to recover broken links. Most of them are based on information annotated in advance with the link. Davis [1] has studied the causes that provoke the existence of broken links and has proposed solutions focussed on collecting information on the links in its creation or modification. The Webvise system [2], integrated with Microsoft software, stores annotations in hypermedia databases external to the web pages. This allows the system to provide a certain degree of capacity to recover integrated broken links. The information is stored when the links are created or modified. Shimada and Futakata

* Supported by project TIN2007-67581-C02-01.

H. Geffner et al. (Eds.): IBERAMIA 2008, LNAI 5290, pp. 302–311, 2008.

[3] designed the Self-Evolving Database (SEDB), which stores only links in a centralized fashion while documents are left in their native formats at their original locations. When a document is missing, the SEDB reorganizes all links formerly connected to the missing document in order to preserve the topology of links.

Nakamizo et al. [4] have developed a software tool that finds new URLs of web pages after pages are moved. The tool outputs a list of web pages sorted by their plausibility of being link authorities. Links contained in the link authority pages are expected to be well-maintained and updated after the linked pages are moved. In this work, a page v is called a link authority of another web page u if (1) v includes the link to u, and (2) if u is moved to u_{new}, the link to u in v is always changed to the link to u_{new}. This system uses a link authority server which collects links to u and then sorts them by plausibility. This plausibility is based on a set of attributes concerning the relations among links and directories.

Thought with a purpose different to repairing broken links, other works 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. McBryan [5] proposed to use the anchor text as a help to the search of web resources. This work describes the tool WWWW intended to locate resources on the Internet. The WWWW program surfs the Internet locating web resources and builds a database of these. Each HTML file found is indexed with the title string used in there. Each URL referenced in an HTML file is also indexed. The system allows searching on document titles, reference hypertext, or within the components of the URL name strings. Chakrabarti et al. [6] have developed an automatic resource compiler which, given a topic that is broad and well-represented on the web, seeks out and returns a list of web resources that it considers the most authoritative for that topic. The system is built on an algorithm that performs a local analysis of both text and links to arrive at a "global consensus" of the best resources for the topic.

Our work differs from previous proposal since it does not rely on any information about the links annotated in advance, and it can be applied to any web page.

Sometimes we can recover a broken link by entering the anchor text as a user query in a search engine. 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 — the text of the web page that contains the link, the page stored in the cache of the search engine, if it exists, the Url, etc.— to the anchor text of the broken link.

In this work we have developed a system to perform this process automatically. 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. The candidate pages are obtained by submitting to a search engine queries composed of terms extracted from different sources. In order to tune the results, the pages recovered in this way are ranked according to relevance measures obtained by applying information retrieval techniques. The resulting list of pages is presented to the user. Figure 1 presents a scheme of the proposed system.

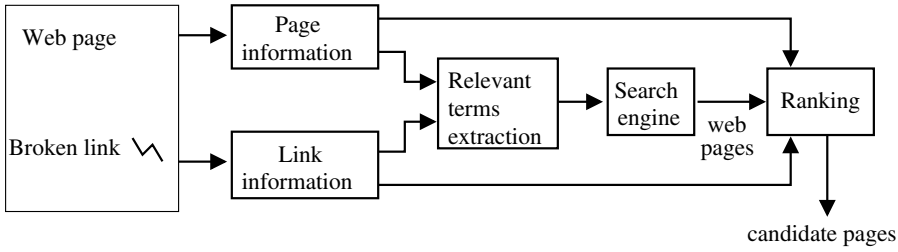

Fig. 1. Scheme of the system for automatic recovering of broken links

This work has begun by analyzing a large number of web pages and their 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. This study has allowed us to extract criteria to determine, for a particular broken link, whether it makes sense to look for candidate pages to recommend to the user, or whether the available information is not enough to attempt the recovering.

The remaining of the paper proceeds as follows: section 2 describes the methodology we have followed to evaluate the suitability of the sources of information considered; section 3 presents the analysis of those sources of information; section 4 is devoted to describe the process to rank the candidate documents; section 5 describes the scheme resulting of the previous analysis, as well as the results of applying it to a set of broken web links. Finally, section 6 draws the main conclusions of this work.

2 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 candidate pages to replace the link. 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.

2.1 Selection of Links to Recover

To carry out the analysis, we take links from pages selected randomly by means of successive requests to *www.randomwebsite.com*, a site that provides random web pages. Certain requisites are imposed to our test pages. We tried to restrict the language to English, considering the following domains: ".com ", ".org ", ".net ", ".gov " and ".edu". Pages with at least 250 words are required for using its text to characterize this page. Moreover, the text will have to contain at least ten terms that are not stop words, that is, words that are so common that they are ignored in information retrieval (e.g. articles, pronouns, etc.) We also demand that the page have at least five potentially *analyzable links*, which means:

- The system analyzes external links, therefore links that point to the same site are discarded.
- The anchor text must neither be empty nor be a number or an URL.
- If the anchor text is only formed of one character and it also coincides with a punctuation mark, this link is discarded.

Pages not fulfilling these requirements are discarded, and the selection process does not finish until one hundred pages are collected, what amounts to having at least 500 links to study. Some preliminary experiments indicated us that it is frequent to find pages in which most of the links are online and others in which most of them are broken. When these pages have many links, they bias the results in some way or another. Because of this we have decided to limit the number of links taken per page to ten. This subset of links is randomly chosen among the analyzable links in the page.

3 Sources of Information

In this section we analyze each source of information considered, extracting statistics of its usefulness for the recovery of links when they are applied separately or combined.

3.1 Anchor Text in a Hyperlink

In many cases the words which compose the anchor text of a hyperlink are the main source of information to identify the pointed page. To verify this theory we have performed a study which can be observed in the Table 1. This table shows the number of cases in which the broken links have been recovered, searching in Google the anchor text in inverted commas.

A combination of different tests to consider that a link has been recovered has been used. First of all it is verified if the Url from the page candidate to replace the link matches the analyzed link (remember that in this analysis the link is not really broken). Nevertheless, we have found some cases in which the recovered page has the same content as that of the pseudobroken link, but different Url. Therefore if the Urls do not match, we verify whether the web page content is the same. We have also found several cases in which the page content is not identical, but they were very similar: there are some small changes like advertisements, dates, etc. For this reason, if the contents are not exactly the same, we apply the vector space model [7], 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. For lower values than this threshold (e.g. 0.8), we sometimes recover different pages. We have measured the number of recovered links according to this threshold. Table 1 shows these results. We can observe that using a similarity threshold of 0.9, 41% of the links are recovered in the top ten results (*Google*). In addition, 66% of the recovered links appear in the first position. These results prove that the anchor text is a good source of information to recover a broken link. Lowering the similarity threshold adds very

Table 1. Results of searching the anchor text in Google in terms on the similarity degree used. First column indicates the similarity degree used. *1st pos.* represents the number of pseudobroken links recovered in the first position from the results of the search engine, and *1-10 pos.* the number of those recovered among the first 10 positions. *N.R.L.* represents the links have not been recovered.

Sim. Degree	1st pos.	1-10 pos.	N.R.L.
0.9	253	380	536
0.8	256	384	529
0.7	258	390	521
0.6	262	403	504
0.5	266	425	478

few additional links to the list of recovered ones. Besides, doing this increases the number of wrong results. For these reasons, we have chosen for the degree of similarity a threshold value of 0.9.

Sometimes the anchor terms are little or no descriptive at all. 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.

In this work we have chosen to 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. There exist several software solutions for this task, such as *LingPipe*, *Gate*, *FreeLing*, etc. There also exist multiple resources, like *gazetteers*. But none of these solutions have provide precise results working with the anchors, perhaps because we are working in a wide domain. In addition, the size of the anchor texts is too small for the kind of analysis usually performed by these systems.

Accordingly, we have decided to use the opposite strategy. Instead of finding named entities, we have chosen to compile a set of dictionaries to discard the common words and numbers, assuming that the rest of words are named entities. Although we have found some false negatives, as for example the company "Apple", we have obtained better results using this technique.

Table 2 shows the number of pseudobroken links recovered depending on the presence of named entities in the anchors, and on the number of anchor terms. We can see that when the anchor does not contain any named entity, the number of links that are not recovered is much higher than the number of the recovered ones, whereas both quantities are similar when there exist named entities. This proves that the presence of any named entity in the anchor favors the recovery of the link. Another 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[1]. When the anchor contains named entities, even if there is only one, the number

[1] These few cases are usually Url domains with a common name, e.g. the anchor "Flock" has allowed recovering www.flock.com, the anchor "moo" the Url www.moo.com/flicker, etc.

Table 2. Analysis of not recovered (*N.R.L.*) and recovered links (*R.L.*) according to the type of anchor — with (*Named E.*) and without (*No Named E.*) named entities— and to the number of anchor terms. *4+* refers to anchors with four or more terms.

	Type of anchor			
	Named E.		No Named E.	
Terms	N. R. L.	R. L.	N. R. L.	R. L.
1	102	67	145	7
2	52	75	91	49
3	29	29	27	45
4+	57	61	33	47
total	240	232	296	148

of retrieved cases is significant. Another fact that we can observe is that from two terms on, the number of anchor terms does not represent a big change in the results.

3.2 The Page Text

The most frequent terms of a web page 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 the person to whom the page corresponds. However, in many cases, the forename and surname do not identify a person in a unique way. For example, if we search in Google "Juan Martínez", we obtain a huge amount of results (99.900 aprox. at the time that this paper was wrote). The first result of the search engine which corresponds to Juan Martínez Romo appears in the tenth position. However, if we expand the query using some term present at his web page, such as "web search", then his personal web page goes up to the first position. This example shows how useful is using a suitable choice of terms.

We have applied classical information retrieval techniques to extract the most representative terms from a page. After eliminating the stop words, we generate a term list ranked by frequencies. 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.

In Table 3 we can observe that the expansion considerably increases the number of links recovered in the first ten positions. In spite of this, the number of recovered links in the first position is reduced.

Table 4 shows the number of cases in which the expansion has improved or worsened the results. We can see that, although the number of cases in which the expansion improves the results is quite higher (almost twice: 90 against 52), the number of cases in which it get worse is not negligible. Accordingly, we think that the most suitable mechanism is to apply both recovery ways, later ranking the whole set of results to present the user the most important ones in the first places.

Analyzing the cases in which it becomes possible to recover the correct page with and without named entities and according to the number of terms of the

Table 3. Analysis of the number of retrieved documents in the first position (*1st pos.*), those retrieved among the first 10 positions (*1-10 pos.*) and the links that have not been recovered (*N.R.L.*), according to whether we use query expansion (*EXP*) or not (*No EXP*).

Analysis	1st pos.	1-10 pos.	N.R.L.
No EXP	253	380	536
EXP	213	418	498

Table 4. Number of cases in which the query expansion improves and worsens the results

Expansion	N of Cases
Improvements	90
Worsenings	52

Table 5. Number of not recovered (*N.R.L.*) and recovered links (*R.L.*) according to the type of anchor, with (*Named E.*) and without (*No Named E.*) named entities, and to the number of anchor terms, when the query expansion method is applied. *4+ term.* refers to anchors with four or more terms.

	Type of anchor			
	Named E.		No Named E.	
Terms	N. R. L.	R. L.	N. R. L.	R. L.
1	104	65	127	25
2	55	72	70	70
3	30	28	22	50
4+	59	59	31	49
total	248	224	250	194

anchor (Table 5) we observe that the results are better than without expansion. However, they present the same trends as in the case without expansion: the worst result corresponds to the case with an only term and without named entities, and, in general results are better if there are named entities. Nevertheless, with the current method (query expansion) the number of recovered links, when the anchor consists of just a term and it is not a named entity, is 25. This last value can be considered as a significant quantity. This suggests trying to recover using query expansion in this case too, as long as it is possible to validate the obtained results. This validation is explained in section 4.

4 Ranking the Recommended 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

Table 6. Occurrences of the best candidate page in the elaborated ranking, selecting the N best candidates according to the similarity with the reference page: cache or parent page

Occurrences of the Best Candidate							
First N selected docs	**10**	**20**	**30**	**50**	**80**	**100**	**110**
Cache page	301	305	306	307	310	312	313
Parent page	47	105	132	191	263	305	313

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 page pointed to by the broken link saved in the search engine cache, in this case *Google*. 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 parent page.

Once again we have applied the vector space model [7] to study the similarity between the analyzed page and its broken links. With this technique, we calculate the similarity either with the cache page or with the parent page. Table 6 shows the obtained results ranked by similarity with the cache and the parent page. In the first case, most of the correct retrieved documents appear between the first ten documents, therefore if we can retrieve the cache page, we will be able to do very trustworthy recommendations. In the second case, using the similarity with the parent page, the order of the results is worse. Thus, we will resort to this information only if we can not get the cache page.

5 Strategy 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 following recovery process. First of all, it is checked whether the anchor number of terms is just one and whether it does not contain named entities. If both features are found, the retrieval is only attempted provided the link of the missing page appears in the cache, 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. If the page is in the cache, then the recovery is performed, expanding the query (anchor terms) with extracted terms from the parent page. Then the results are ranked and only if any of them is sufficiently similar to the cache content, 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

Table 7. Number of recovered links (best candidate which content is very similar to the missing page) according to his cache similarity, between N first documents ranked by similarity with the parent page. The total number of broken links investigated were twenty five.

First N docs.	R.L.
1-10	12
10-20	7
20-50	6

the parent page. After that, all documents are grouped and ranked according to the cache page if it is available in Google, or according to the parent page otherwise.

We have applied this strategy to links that are really broken, but we have only used those that were present in the Google cache. The reason is that only in this case we can evaluate the results. Table 7 shows the quantity of recovered links (best candidate whose content is very similar to the missing page) ranking the results by means of the similarity with the parent page (the page cache is only used to measure relevance). 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. We can observe that, even if we are not using the cache similarity and we rank with the similarity with the parent page, the system is able to provide useful replacements documents between the first 10 positions in 48% of the cases, and between the 20 first ones in 76% of the cases.

6 Conclusions and Future Work

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 for expanding the queries. In this way, we reduce the ambiguity that would entail the limited quantity of anchor terms. This study has showed that the results are better when the query is expanded than when using only the anchor terms. However, since there are cases in which the expansion worsens the recovery results, we have decided to combine both methods, later sorting the documents obtained by relevance, to present to the user the best candidate pages at the beginning. 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 the top ten of the results in 48% cases, and between the top 20 in 76%. At this moment we work in analyzing other sources of information that can be useful for the retrieval, as the Urls or the pages that point to the page which contains the broken link.

References

1. Davis, H.: Hypertext link integrity. In: ACM Computing Surveys Electronic Symposium on Hypertext and Hypermedia, vol. 31(4) (2000)
2. Grønbæk, K., Sloth, L., Ørbæk, P.: Webvise: Browser and proxy support for open hypermedia structuring mechanisms on the world wide web. Computer Networks 31(11-16), 1331–1345 (1999)
3. 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, Washington, DC, USA, vol. 2, p. 226. IEEE Computer Society, Los Alamitos (1998)
4. 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)
5. McBryan, O.A.: GENVL and WWWW: Tools for Taming the Web. In: Nierstarsz, O. (ed.) Proceedings of the first International World Wide Web Conference, CERN, Geneva, p. 15 (1994)
6. Chakrabarti, S., Dom, B., Gibson, D., Kleinberg, J., Raghavan, P., Rajagopalan, S.: Automatic resource list compilation by analyzing hyperlink structure and associated text. In: Proceedings of the 7th International World Wide Web Conference (1998)
7. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval. Cambridge University Press, Cambridge (2008)