Why are Moved Web Pages Difficult to Find? The WISH Approach

Atsuyuki Morishima
University of Tsukuba
1-2 Kasuga, Tsukuba
Ibaraki, Japan
mori@slis.tsukuba.ac.jp

Akiyoshi Nakamizo*
Shibaura Institute of Technology
Tokyo, Japan
mizo@sic.shibaua-it.ac.jp

Toshinari Iida*
University of Tsukuba
1-2 Kasuga, Tsukuba
Ibaraki, Japan
toshi@slis.tsukuba.ac.jp

Shigeo Sugimoto
University of Tsukuba
1-2 Kasuga, Tsukuba
Ibaraki, Japan
sugimoto@slis.tsukuba.ac.jp

Hiroyuki Kitagawa
University of Tsukuba
1-1-1 Tennohdai, Tsukuba
Ibaraki, Japan
kitagawa@cs.tsukuba.ac.jp

ABSTRACT

This paper addresses the problem of finding new locations of moved Web pages. We discuss why the content-based approach has a limitation in solving the problem and why it is important to exploit the knowledge on where to search for the pages.

Categories and Subject Descriptors

H.3.3 [**Information Systems**]: Information Search and Retrieval

General Terms

Algorithms, Experimentation

Keywords

Broken links, integrity management

1. INTRODUCTION

This paper addresses the problem of automatic correction of broken Web links focusing, in particular, on links broken by the relocation of Web pages, i.e., finding new locations of moved Web pages, and it reveals how the problem is different from typical information retrieval problems. Although the problem of broken links has been considered a serious problem [1][7] and some studies have addressed possible approaches to link correction [5], software support for the correction of broken links is still in its infancy.

Researchers have been tackled the problem for years, by taking various *content-based approaches*. Peridot is a software tool developed by IBM based on their patents [2][3] and has a function to tackle the problem: It automatically tries to find new links for broken links. It computes the likelihood of pages being the destination based on *fingerprints*, which are information on the *contents* of Web pages. Phelps and Wilensky proposed a *lexical signature* [9], which is a small set of keywords chosen to effectively identify the intended Web page when it is submitted to index servers. The approaches taken by Peridot and lexical signatures are similar in that their emphasis is on how to use *contents* of Web pages to find destinations. In content-based approaches, information on the content of the destination page *must* be stored in databases or index servers in advance.

Park et al. [4] compared the lexical signature and other content-based methods like TF-IDF in experiments and showed that most of content-based methods were effective for 60 to 70 percent of pages; Many search engines returned the original page as the 1st ranked result. Note, however, that the results were obtained under the following conditions. (1) all the pages were indexed, and (2) the contents were not changed (in addition, the pages were not moved).

In this paper, we argue that the application of the content-based methods alone is not sufficient to find new locations of moved Web pages, and explain why we need to exploit the knowledge on where to search for the pages in solving the problem.

2. FINDING NEW LOCATIONS OF MOVED WEB PAGES: THE REALITY

We conducted a relatively large experiment to evaluate different approaches to finding new locations of moved Web pages [8]. The experimental results revealed that there is a limitation of the index-server (i.e., content-based) approach even if we use the index-servers that would have *the largest amount of Web indices at present* (Google, Yahoo, and MSN).

In the experiment, we analyzed 858 broken links, which we found from the 127,109 *out-going* links appearing in nine university domains. We identified 259 links of the 858 broken links, which had been caused by page movements. New

*Currently, Hitachi, Ltd.

WWW 2009, April 20–24, 2009, Madrid, Spain.
ACM 978-1-60558-487-4/09/04.

locations for the 259 links were identified after we carefully searched for their destinations manually. This number is quite interesting, since almost one-third of broken links were caused by page movements.

The findings in our experiment are that in real situations, not many Web pages are immediately indexed after they are moved, and that the index-server approach is not necessarily effective because it is impossible to identify the final destination until the page is moved. With the help of the cached pages we got in advance, we extracted keywords to search for the new locations of the Web pages, and submitted them to the index servers. The result is that we could find only 27.8% of the new locations at best, partly because some of the pages were not indexed. More interestingly, there were pages whose contents had been drastically changed. For example, some of the moved Web pages were changed from text pages to Flash pages. The result suggests that we need alternatives or complementary approaches to the content-based approach. In fact, Phelps and Wilensky themselves mentioned some cases where the content-based approach is not effective. They include (1) non-indexed documents, and (2) resources with highly variable content. The results in [4] supported the argument by showing that the content-based method became much less effective when the contents were changed.

3. WHY THE LOCATION-FACTORS ARE IMPORTANT: THE WISH APPROACH

A natural question might be: how could we find the new locations for the 259 moved Web pages, although the index servers could not find them? The reason is that there is a *bias* about the place the destination page is likely to be located. Consider what a person might do if a broken link is encountered: The person may check if the root page of the web site is still there and then follow links to find the page. Or he may dredge up from memory other pages that had links to the same page to check if their links have been updated. In other words, people seem to know *where* the destination page is likely to be located. This suggests that there is a possibility that we can develop a system to find moved pages efficiently exploiting the bias.

Based on the observation, we argue that we should exploit the knowledge on *where* the destination page is likely to be located, and propose to develop an effective set of heuristics to imitate the process of people searching for new locations of moved web pages. The introduction of the new approach affects the problem of finding new links in the following two ways: (1) It introduces another factor beside the content of pages, and (2) it reduces the size of the search space for new links.

Then, the challenge is to find such a set of heuristics. We are experimenting with different heuristics using the PageChaser, which is a tool we developed in our WISH (Web Integrity management by Self-Healing mechanisms). Most of the heuristics implemented are about where we should search for new links. For example, PageChaser explores Web sites to which PageChaser knows some related pages of the target page have already been moved. In particular, PageChaser introduces a novel concept called *link authority*. A link authority is a special kind of Web page; a link authority p of another page q is a web page in which any link to q contained in p is always kept up-to-date. For example, the official "links to departments" page of a university is a link authority for its department pages. PageChaser provides a crawler-based solution based on the heuristics. From our experience [8], the set of heuristics can produce dramatically better results (could find new locations for more than 70% of the broken links) without the need for an exhaustive search of the entire Web.

4. SUMMARY

This paper argue that the content-based approach alone is not sufficient to find new locations of moved Web pages, and that using the knowledge on where to search for the new locations (introduction of the location-oriented heuristics) to imitate the people's process of searching for new locations, is promising.

5. ACKNOWLEDGMENTS

We would like to thank Prof. Atsushi Toshimori, Tetsuo Sakaguchi, and Mitsuharu Nagamori for their valuable comments. This research was partially supported by the Grant-in-Aid for Scientific Research (#19024006, #19300081, #20800076) from the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan.

6. REFERENCES

[1] H. Ashman, H. Davis: Panel Missing the 404: link integrity on the World Wide Web. Computer Networks 30(1-7): 761-762 (1998)

[2] M. Beynon, A. Flegg: Hypertext Request Integrity and User Experience. US Patent Application Publication, US 2004/0267726 A1, Dec, 2004.

[3] M. Beynon, A. Flegg: Guaranteeing Hypertext Link Integrity. US Patent Application Publication, US 2005/0021997 A1, Jan. 2005.

[4] S. Park, D. M. Pennock, C. L. Giles, R. Krovetz: Analysis of lexical signatures for improving information persistence on the World Wide Web. ACM Trans. Inf. Syst. 22(4): 540-572 (2004)

[5] H. C. Davis: Hypertext link integrity. ACM Comput. Surv. 31(4es): 28 (1999)

[6] Katsumi Tanaka, N. Nishikawa, S. Hirayama, K. Nanba: Query Pairs as Hypertext Links. ICDE 1991: 456-463.

[7] GVU Center, College of Computing Georgia Institute of Technology. GVU's 10th WWW User Survey. http://www.gvu.gatech.edu/user_surveys/survey-1998-10/.

[8] A. Morishima, et al. Automatic Correction of Broken Web Links. Technical Report, University of Tsukuba.

[9] Thomas A. Phelps, Robert Wilensky: Robust Hyperlinks: Cheap, Everywhere, Now. DDEP/PODDP 2000: 28-43