Rediscovering Missing Web Pages Using Link Neighborhood Lexical Signatures

Martin Klein
Dept of Computer Science
Old Dominion University
Norfolk, VA, 23529
mklein@cs.odu.edu

Jeb Ware
Dept of Computer Science
Old Dominion University
Norfolk, VA, 23529
jware@cs.odu.edu

Michael L. Nelson
Dept of Computer Science
Old Dominion University
Norfolk, VA, 23529
mln@cs.odu.edu

ABSTRACT

For discovering the new URI of a missing web page, lexical signatures, which consist of a small number of words chosen to represent the "aboutness" of a page, have been previously proposed. However, prior methods relied on computing the lexical signature before the page was lost, or using cached or archived versions of the page to calculate a lexical signature. We demonstrate a system of constructing a lexical signature for a page from its link neighborhood, that is the "backlinks", or pages that link to the missing page. After testing various methods, we show that one can construct a lexical signature for a missing web page using only ten backlink pages. Further, we show that only the first level of backlinks are useful in this effort. The text that the backlinks use to point to the missing page is used as input for the creation of a four-word lexical signature. That lexical signature is shown to successfully find the target URI in over half of the test cases.

Categories and Subject Descriptors

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

General Terms

Measurement, Performance, Design

Keywords

Web Page Discovery, Preservation, Link Neighborhood

1. INTRODUCTION

At times during a web browsing session, a user's navigation is interrupted by broken links which point to nonexistent pages. In some cases, these pages are not lost forever, but can be accessed at a different address. Lexical signatures, which consist of a small number of words chosen to represent the "aboutness" of a page, have been proposed as a way to discover the new URI of a page. Typically a lexical signature of a web page consists of its top n terms ranked in decreasing order of their TF-IDF value. Term frequency (TF) represents the commonality of a term on the page and inverse document frequency (IDF) indicates the rareness of that term in the document's corpus. For example, a lexical signature for a page that sells instruments for environmental measurement might be *humidity, measurement, machine*. Prior research [7, 9, 10] has shown that using those words as a search engine query can yield the same (or similar) page at its new URI. However, previous methods relied on computing the lexical signature a priori meaning before the page was lost, or using cached or archived versions of the page to calculate its lexical signature. If it had not been previously calculated, and no archived or cached copies of the page were available, these methods are unusable.

To overcome this limitation, we performed an experiment to evaluate constructing lexical signatures from link neighborhoods. Since pages tend to link to related pages, our intuition was that the link neighborhoods contain enough of the "aboutness" of the targeted page to allow a lexical signature to be created. Therefore our experiment is focused on calculating a lexical signature for a URI when no representation of that URI can be retrieved from a search engine's cache or archive. We constructed link neighborhoods by querying a search engine for listings of backlinks and tested several methods of calculating lexical signatures from those link neighborhoods to find the most effective signatures.

We examined the effects of lexical signature size, backlink depth, backlink ranking as well as the radius within a backlink page from which the lexical signature terms are drawn.

2. RELATED WORK

Phelps and Wilensky [10] proposed calculating the lexical signature of a target page, and embedding that lexical signature into the link URIs to make the referenced page easier to find. Their method relied on a 5-term lexical signature being calculated at the time the link was created, and included in the link URI. This placed the burden of preparing for future recovery on the content creator or administrator; if the creator did not calculate the lexical signature in advance, the user would be unable to use this method to attempt to rediscover the page. In addition, web browsers would have to be modified to use the lexical signature in the URI to attempt to rediscover the page.

Park et al. [9] expanded on the work of Phelps and Wilensky by analyzing eight different formulas for calculating a lexical signature. They tested Phelps and Wilensky's orig-

JCDL'11, June 13–17, 2011, Ottawa, Canada.

inal TFIDF variant, a simpler TFIDF, plain TF, and plain DF. In addition they tested several "hybrid" formulas in which the terms were calculated with different formulas, as potential ways to find relevant documents when the original document could not be found. They found that TFIDF was the best among the "basic" formulas, though some of the hybrid formulas performed better in certain use cases. They also used five terms for lack of an empirical study on lexical signature size, and noted that the effect of the size was a topic for future research.

Klein and Nelson [7] proposed a method to use archived or cached versions of a missing page to calculate its lexical signature. They showed that using a 5-term or a 7-term lexical signature as a search engine query would produce the best results for rediscovering that page. Five terms did best at finding the URI as the top result, whereas seven terms performed best at finding the URI in the top ten results.

Henzinger et al. [5] used lexical signatures derived from newscast transcripts to find articles related to the newscast in real-time. Their input, rather than being a static web page, was a constantly-flowing stream of text from the transcript. Their method took into account the temporal locality of terms, that is words that were spoken close together in the broadcast, to attempt to compute lexical signatures that would be relevant to a single story each, rather than spanning across subsequent stories. Their observations showed that, far from the five terms used in prior studies, a two-term lexical signature worked best in this application.

Craswell et al. [3] showed the effectiveness of anchor text in describing a resource. They demonstrated that for a specific user need (the site-finding problem) anchor text of backlinks provided a more effective way of ranking documents than did the content of the target page itself.

Sugiyama et al. [11] proposed enhancing the feature vector of a web page by including its link neighborhood. That is, they proposed that a search engine could more accurately describe the contents of a page by including information from both in-links (backlinks) and out-links. They tested up to third-level in- and out-links, and found that only links up to the second level were helpful. In some of the methods they tested, only the first-level links were helpful.

Fujii et al. [4] explored the correlation between anchor text and page titles. They showed that the collective nature of anchor texts, since they are created by many people, adds a significant value. Anchor texts are created by a similar thought process to queries, and as such will use similar words to describe a topic. Since links can be made by many authors, they will use their own word preferences, which means that anchors can provide synonyms that the original page's author might not use. They even showed that anchor texts, since they might be written in different languages, might be used to provide a bilingual corpus for machine learning of natural language translation.

3. METHODOLOGY

We obtained 309 URIs for our experiment from the same corpus used in Klein and Nelson 2008 [7]. For each URI, we queried the Yahoo! BOSS[1] Application Programming Interface (API) to determine the pages that link to the URI ("backlinks"). We chose the Yahoo! BOSS API because it was previously shown to give more complete backlink results than other search engines [8]. We refer to the order in which these backlinks are returned as "backlink rank". By obtaining the backlinks of the backlinks we created a directed graph of depth 2. Figure 1 shows an example of such a link neighborhood. The page on the right with the vertical lines represents the target page, the page that a user linked to but is no longer available. Using Yahoo!, we obtain (in this example) three pages that link to the target page. These are the first-level backlinks, represented in the center with horizontal lines. For each first-level backlink, we obtain its backlinks, represented with crossing lines as the second-level backlinks. In this manner we retrieved 28,325 first-level and 306,700 second-level backlink pages.

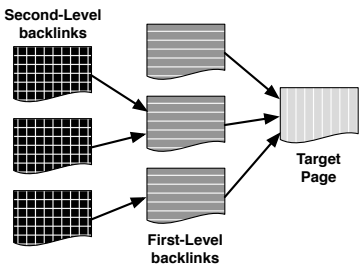

Figure 1: Example Link Neighborhood

3.1 Pre-processing Link Neighborhood

In order to guard against noisy inputs we applied five filters to the backlink page representations before we calculated lexical signatures. Those filters were based on: content language, file type, file size, HTTP return code, and the presence of "soft 404s". Since we used an American search engine which is likely to have coverage biased towards English-language pages [12] we made an attempt to discard non-English pages before lexical signature calculation. We used a Perl module to obtain percentage guesses for each page's language and dismissed less than 8% of all pages. We discarded less than 4% of all pages due to their non-HTML representations. In cases where the server described the file type in the HTTP headers, this information was trusted. In other cases, the Unix 'file' command-line program was used to guess the file type. Similar to the filter applied by Park et al. [9] and several subsequent studies we discarded any pages that contained less than fifty terms after rendering to ensure enough input for a robust lexical signature. This filter accounted for dismissing less than 7% of all pages. We experienced a variety of errors while downloading all web pages. Fortunately almost 97% of pages returned the HTTP response code 200 which means success. We made a total of five attempts to download the remaining pages. If none of the attempts resulted in success, those pages were dismissed. Pages that terminate with an HTTP 200 status code (possibly after multiple redirects) while returning an error message in the human-readable version of the page like "your page could not be found" are known as "soft 404s". Since these pages are not "about" the same things as the pages that they link to, they were discarded (less than 1%). To determine if a page was a soft 404, we used a subset of the method laid out by Bar-Yossef et al. [2].

For more details about all applied filters as well as fur-

[1] `http://developer.yahoo.com/search/boss/`

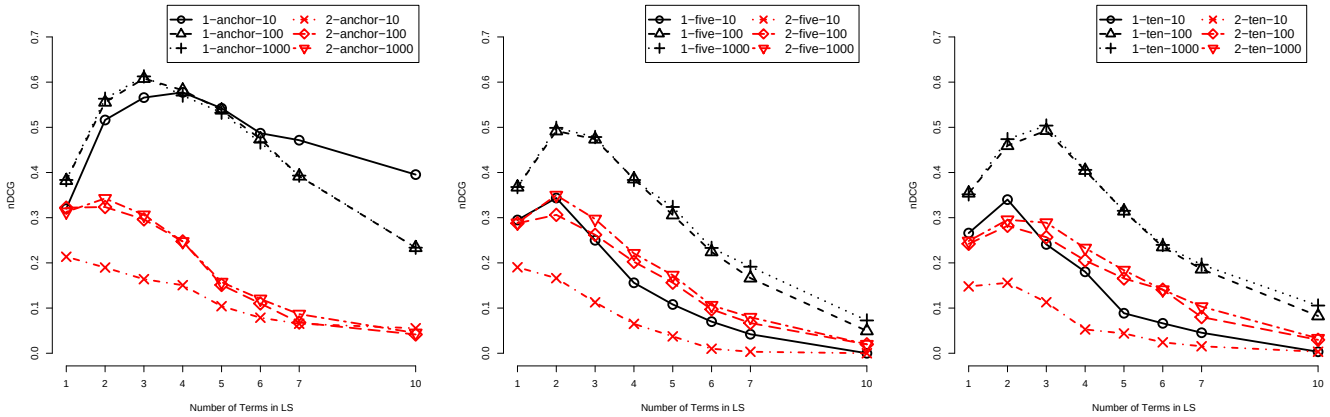

Figure 2: First- and Second-Level Backlinks, All Radii and Backlink Ranks (shown as level-radius-ranks)

ther statistics we refer to the technical report version of this paper [13].

3.2 Lexical Signature Calculation

As mentioned in the Introduction, we sought to determine the effects of lexical signature size, backlink depth, backlink ranking, as well as the radius within a backlink page from which terms for the lexical signature should be drawn. For every possible combination for each of these factors we computed the TFIDF value of every term in the appropriate section(s) of the appropriate pages. The terms with the highest TFIDF value were taken as the lexical signature.

Backlink Depth: The two options for depth were (1) to use only the first-level backlinks, those that link directly to the target page and (2) to use first and second level backlinks. First-level backlinks being closer to the target page, they might result in a lexical signature that more accurately describes the missing page. However in cases where few first-level backlinks exist, second-level backlinks might provide more information, leading to a better lexical signature.

Backlink Ranking: The backlinks returned from the Yahoo! BOSS API are ordered. To test whether this ranking was helpful we tested the following three possibilities: 1) using only the top ten backlinks, 2) using the top hundred backlinks, and 3) using the top thousand backlinks. If fewer backlinks existed than allowed by the limit then all available backlinks were used. If the rankings in backlink results from the BOSS API were helpful, then using only the top backlinks would provide a better lexical signature. If not, then using as many backlinks as possible might provide the better lexical signature by including more data.

Radius: We considered four possibilities for the radius within the backlink page from which lexical signatures would be drawn. Lexical signatures are typically drawn from entire pages, and this was our first possibility. However, since a particular section of a page can be about a different topic than a page as a whole, we tested whether using only the relevant portions of a page would produce a better lexical signature. To find the "relevant" portion of a backlink page we used the link from the page to the target URI as a centerpoint and captured the 'paragraph' of context around the link. Hence the second option considered the anchor text plus the preceding five words and the following five words. Option three included the link plus ten words on each side. The fourth option used only the anchor text itself. In cases where a given backlink page included multiple links to the same target URI, the text around every link was included. In cases where the link to the target URI could not be found, the backlink page was not included in the calculations.

Lexical Signature Size: Most research using lexical signatures recommends using 5- or 7-term lexical signatures. However, given that the lexical signatures in this experiment were being derived from a link neighborhood instead of the target page itself, the applicability of those standards was tested. We stored the ten terms with the highest TFIDF value and queried lexical signatures of sizes one, two, three, four, five, six, seven, and ten.

3.3 Scoring of Results

Since users tend not to look past the first few results for a query [1, 6], a scoring system rewarding results at the top of the result set and penalizing results on the lower end was necessary. We used normalized Discounted Cumulative Gain (nDCG) for this purpose with a relevance score of 1 for an exact match of the target URI, and 0 otherwise. We checked only the first 1000 results and if the target URI was not found we assigned a nDCG value of 0, corresponding to an infinitely deep position in the result set.

4. RESULTS AND DISCUSSION

Figure 2 shows average scores of methods based on anchor text, anchor text ±5 words and anchor text ±10 words using the first-level (black) and second-level backlinks (red). Since the method based on using the whole page performed exceptionally poorly and its results do not change the overall outcome we decided to not include it in the graph. The x-axis is the number of terms included in the lexical signature, and the y-axis is the mean nDCG.

Note the dramatic decline in every case when second-level backlinks are included. This shows that second-level backlinks' relation to the target page is not tight enough to be useful in describing the target page. As such, our best-performing method includes only first-level backlinks.

We started with the assumption that some parts of a backlink page would use terms that are more closely related to the target page, and that the most relevant terms would be

in or near the link to the target URI. We see in the above figure that by far the best results arise from using only the terms in the anchor text itself to calculate the lexical signature. The anchor text ±5 words or ±10 words performed similar to each other. Each step taken away from the anchor text, by broadening the radius to include words around the anchor or the entire page, yields poorer and poorer results.

The figure also shows the three possibilities for backlink ranks: using only the top ten, top 100, or top 1000 backlinks. Note that using 100 and 1000 backlinks (and anchor text) results in the highest overall nDCG value of 0.61 obtained with 3-term lexical signatures. This accounts for more than 58% of all URIs returned as the top search result. The corresponding nDCG value using the top ten backlinks only is 0.57. Considering this marginal delta in nDCG and the huge implied cost to acquire ten or one hundred times as many pages and generate a lexical signature based on an accordingly larger bucket of words, we consider using only the top ten backlinks as the better tradeoff. The "return on investment" is better when sacrificing an nDCG drop of only 0.04. So while we do not know how Yahoo! determines its ranking, we do know (considering all costs) that we are better off using only the top ten backlinks of a URI.

We can further see that the overall best performance is obtained using 3-term lexical signatures. However, with the above reasoning meaning to use the top ten backlinks only, the best-performing lexical signature is four terms in length. Using ten backlinks, 4-term lexical signatures have an nDCG value of 0.58 and almost 56% of all URIs are returned top ranked. Using more or fewer terms yields poorer results. This result is noteworthy in that most other implementations of lexical signatures found that the best lexical signature size is five or seven terms [7]. The reason for the disparity is the source of the terms that make up the lexical signature. In this method, the terms are drawn not from the target page itself, but from pages that link to it, which are likely to be "related". Using five or seven terms drawn from the backlink pages is likely to over-specify the backlink pages and their specific focus, rather than the content of the target page. By using fewer terms, we decrease the risk of including a term in the lexical signature that does not appear in the target page.

For more details about all results we again refer to our technical report [13].

5. CONCLUSIONS

In this paper, we showed that lexical signatures calculated from the backlink neighborhood of a web page can be used to re-discover that page. This method can be used when no copies of the missing page exist, and the page was not analyzed before it went offline.

We found that only the first-level backlinks were related closely enough to the missing page to be helpful in creating a lexical signature. Including second-level backlinks introduced too much noise into the signatures. The most relevant text in a backlink page is the text that is used to link to the missing page, and this "anchor text" provides enough terms to develop a viable lexical signature for the missing page. Our results using the top ten, 100 and 1000 backlinks (with anchor text) are very similar. Hence the ranking of backlinks for a URI is not essential to us in determining a lexical signature for the missing URI. However, due to greater costs for 100 and 1000 backlinks we recommend using the top ten only. It remains for future work to determine whether there is an optimal value for backlinks since our three cutoffs were chosen arbitrarily. The best performing lexical signature derived from the top 10 backlinks contains four terms. Using the top 100 and 1000 backlinks three term lexical signatures perform slightly better. Any fewer terms and the signature is too vague and with any more terms the signature runs the risk of including terms not present in the target page, which likely excludes the target URI from the results.

6. ACKNOWLEDGEMENTS

This work was supported in part by the Library of Congress.

7. REFERENCES

[1] E. Agichtein and Z. Zheng. Identifying "Best Bet" Web Search Results by Mining Past User Behavior. In *Proceedings of KDD '06*, pages 902–908, 2006.

[2] Z. Bar-Yossef, A. Z. Broder, R. Kumar, and A. Tomkins. Sic Transit Gloria Telae: Towards an Understanding of the Web's Decay. *Proceedings of WWW '04*, pages 328–337, 2004.

[3] N. Craswell, D. Hawking, and S. Robertson. Effective site finding using link anchor information. In *Proceedings of SIGIR '01*, pages 250–257, 2001.

[4] A. Fujii, K. Itou, T. Akiba, and T. Ishikawa. Exploiting anchor text for the navigational web retrieval at ntcir-5. In *Proceedings of NTCIR-5 '05*, Tokyo, Japan, December 2005.

[5] M. Henzinger, B.-W. Chang, B. Milch, and S. Brin. Query-Free News Search. *Proceedings of WWW '03*, pages 1–10, 2003.

[6] T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay. Accurately Interpreting Clickthrough Data as Implicit Feedback. In *Proceedings of SIGIR '05*, pages 154–161, 2005.

[7] M. Klein and M. L. Nelson. Revisiting Lexical Signatures to (Re-)Discover Web Pages. *Proceedings of ECDL 2008*, pages 371–382, 2008.

[8] F. McCown and M. L. Nelson. Agreeing to disagree: search engines and their public interfaces. In *Proceedings of JCDL '07*, pages 309–318, 2007.

[9] S.-T. Park, D. M. Pennock, C. L. Giles, and R. Krovetz. Analysis of Lexical Signatures for Improving Information Persistence on the World Wide Web. *ACM Transactions on Information Systems*, 22(4):540–572, 2004.

[10] T. A. Phelps and R. Wilensky. Robust Hyperlinks Cost Just Five Words Each. Technical Report UCB/CSD-00-1091, EECS Department, University of California, Berkeley, 2000.

[11] K. Sugiyama, K. Hatano, M. Yoshikawa, and S. Uemura. Refinement of TF-IDF Schemes for Web Pages using their Hyperlinked Neighboring Pages. *Proceedings of HT '03*, pages 198–207, 2003.

[12] L. Vaughan and M. Thelwall. Search Engine Coverage Bias: Evidence and Possible Causes. *Information Processing and Management*, 40(4):693 – 707, 2004.

[13] J. Ware, M. Klein, and M. L. Nelson. An Evaluation of Link Neighborhood Lexical Signatures to Rediscover Missing Web Pages. Technical Report arXiv:1102.0930v1, Old Dominion University, 2011.