2014

Moved But Not Gone: An Evaluation of Real-Time Methods for Discovering Replacement Web Pages

Martin Klein
Los Alamos National Laboratory

Michael L. Nelson
Old Dominion University

Original Publication Citation

Klein, M., & Nelson, M. (2014). Moved but not gone: An evaluation of real-time methods for discovering replacement web pages. *International Journal on Digital Libraries*, *14*(1/2), 17-38. https://doi.org/10.1007/s00799-014-0108-0

Int J Digit Libr (2014) 14:17–38
DOI 10.1007/s00799-014-0108-0

Moved but not gone: an evaluation of real-time methods for discovering replacement web pages

Martin Klein · **Michael L. Nelson**

Received: 27 November 2012 / Revised: 13 January 2014 / Accepted: 13 January 2014 / Published online: 1 February 2014

Abstract Inaccessible Web pages and 404 "Page Not Found" responses are a common Web phenomenon and a detriment to the user's browsing experience. The rediscovery of missing Web pages is, therefore, a relevant research topic in the digital preservation as well as in the Information Retrieval realm. In this article, we bring these two areas together by analyzing four content- and link-based methods to rediscover missing Web pages. We investigate the retrieval performance of the methods individually as well as their combinations and give an insight into how effective these methods are over time. As the main result of this work, we are able to recommend not only the best performing methods but also the sequence in which they should be applied, based on their performance, complexity required to generate them, and evolution over time. Our least complex single method results in a rediscovery rate of almost 70 % of Web pages of our sample dataset based on URIs sampled from the Open Directory Project (DMOZ). By increasing the complexity level and combining three different methods, our results show an increase of the success rate of up to 77 %. The results, based on our sample dataset, indicate that Web pages are often not completely lost but have moved to a different location and "just" need to be rediscovered.

Keywords Missing Web Pages · Web Page Discovery · 404 Error · Web Preservation · Web Archives · Memento

M. Klein (✉)
Research Library, Los Alamos National Laboratory,
Los Alamos, NM, USA
e-mail: mklein@lanl.gov

M. L. Nelson
Computer Science Department, Old Dominion University,
Norfolk, VA, USA
e-mail: mln@cs.odu.edu

1 Introduction

Inaccessible Web pages and 404 "Page Not Found" responses are part of the Web browsing experience. Despite guidance for how to create "Cool URIs" that do not change [1], there are many reasons why URIs or even entire websites break [2]. A 404 response constitutes a detriment to the user's browsing experience but it is our intuition that information on the Web is rarely completely lost, it is just missing. In whole or in part, content often has just moved from one URI to another and, consequently, it becomes an issue of rediscovering it at its new location. In this paper, we propose the use of the following four retrieval methods for this purpose:

1. lexical signatures,
2. Web pages' titles,
3. tags, and
4. link neighborhood lexical signatures.

If these methods perform well, they could (automatically) be applied whenever a user encounters a 404 response and as the result, relevant alternatives to the initially requested Web page could be offered while the user is browsing. Users that are just interested in finding missing pages and digital preservation researchers alike could benefit from such an implementation.

Web archives such as the Internet Archive (IA) provide copies of Web pages dating back as far as 1996. The IA continuously crawls the Web and makes copies of Web pages freely available to the public. The Memento framework [3] utilizes multiple such Web archives to enable "time travel" for the Web. Memento takes a URI and a desired datetime of the past and returns the Web resource's representation as available at the specified time from a Web archive. An old copy of a Web resource is called a Memento. The first two of the here investigated methods, the lexical signature and the title of missing Web pages, are retrieved from the pages' Mementos. The Mementos are obtained using the URI of the missing page in the Memento framework. These methods are, therefore, not applicable if no Mementos of the Web pages are available. Tags and link neighborhood lexical signatures, on the other hand, do not rely on the availability of Mementos of missing pages. They can be obtained and generated with the help of third party indexes such as social annotation services and search engines. The URI of the missing page serves as the basis of these two methods also.

Sometimes, a Memento of a missing page is sufficient for the user's information need, especially for rather static pages, but sometimes one needs to find the current information on the live web.

This article presents the compressed main results of the author's dissertation work [4] on the recovery of missing Web pages in real time. We see this work as the nexus of Information Retrieval (IR) and digital preservation; we are investigating the performance of techniques that can be generated with methods not unknown in the IR world to address a vital Web preservation issue. This work evaluates the individual retrieval performance of all four above mentioned methods as well as the performance of various logical combinations of methods.

The presented experiments were conducted over a period of five years and are based on various different corpora. As shown in [5] and [6], finding a reasonably sized sample set of URIs representing the entire Web is a non-trivial task. Rather than attempting to get an unbiased dataset, we randomly sampled URIs from the Open Directory Project (DMOZ)[1] for most of our corpora. While DMOZ has been used by various researchers in the past, for example [7–9], this choice has several consequences that need to be taken into consideration when evaluating the here presented results.

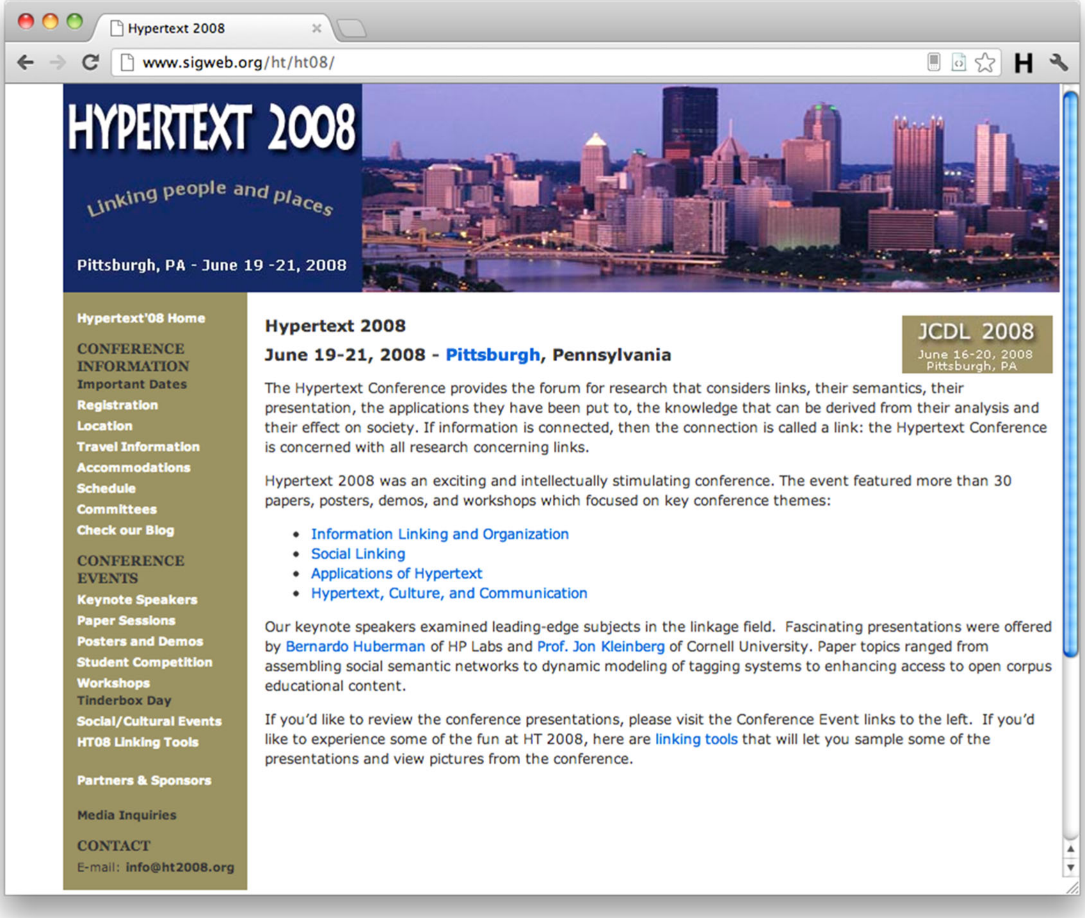
1. DMOZ URIs are usually not missing. However, we are not aware of any available corpora comprising suitably sized set of missing Web pages and hence we measure the performance of our methods as if the URIs were missing. Hence our results somewhat represent an upper bound of what can be expected. Note that during the late phase of the dissertation work, an effort was started to generate a corpus of missing Web pages called "Book of the Dead" [10]. However, this is an ongoing effort and a more detailed report remains for future work.
2. DMOZ is a manually catalogued set of URIs and hence the case can be made that they do not really represent the true nature of the Web, especially when considering what is known as the "deep Web". They do, however, represent a notion of popularity, meaning that this set of URIs corresponds to what people (versus robots processing entire collections) actually consume. In addition, we have seen an unprecedented acceleration in archives and archiving technology, meaning that the presence of copies of Web pages in archives is steadily improving compared to the late 1990s where the IA was the only reliable source of Mementos. Many more public web archives are now available, for example, Archive.is. Also, the IA had a 6–12 months quarantine period in which they had a Memento of a Web page but did not make it available in their index. This quarantine is now gone which means that they are ingesting Mementos into their index as soon as the pages are crawled. All that is a vast improvement of the Web page archiving landscape.
3. It has recently been shown in [11] that archives are not necessarily independent of factors such as popularity. In particular, DMOZ is often used as a seed list for crawlers and archives and as such, it does represent a sort of best case for archival coverage. The answer to the question of how much of the Web is archived in [11] depends heavily on the sample set. URIs sampled from DMOZ and Delicious, for example, show a much higher rate of copies in Web archives (79 and 68 %, respectively) than URIs sampled from other sources such as Bitly (16 %).

A good example for the motivation of this work is the website of the Hypertext Conference in 2008. The original URI http://ht2008.org is not accessible anymore and returns a 404 error today. However, a simple query to a search engine returns the new location of the original content at the new URI http://www.sigweb.org/ht/ht08. Figure 1 shows a screen shot of the content at its new location.

The title of the Hypertext 2008 Web page is *Hypertext 2008* and its lexical signature is *Hypertext Conference presentations Linking SIGWEB logo conference*. Both values are obtained from the latest available Memento of the page provided by the IA. If used as the query string for a search engine, both the lexical signature and the title-based method return the new location of the content in the top three results.

The remainder of the article is structured as follows: Sect. 2 gives an overview of related research that motivated this work. Section 3 introduces the notion of a lexical signature and briefly recaps how they can be generated and applied to Web pages while covering the most relevant related research on the topic. It describes experiments on the performance of lexical signatures in terms of their length, the index they were generated from, and their age. Section 4 details our experiments on the performance of Web pages' titles and the combination of titles and lexical signatures. It also provides insights into our study of the title evolution over time. Section 5 describes the experiments on the performance of tags by themselves and in combination with other methods. It also introduces our notion of "Ghost Tags". Section 6 includes

[1] http://dmoz.org.



HYPERTEXT 2008

Linking people and places

Pittsburgh, PA - June 19 -21, 2008

Hypertext'08 Home

CONFERENCE INFORMATION
Important Dates
Registration
Location
Travel Information
Accommodations
Schedule
Committees
Check our Blog

CONFERENCE EVENTS
Keynote Speakers
Paper Sessions
Posters and Demos
Student Competition
Workshops
Tinderbox Day
Social/Cultural Events
HT08 Linking Tools

Partners & Sponsors

Media Inquiries

CONTACT
E-mail: info@ht2008.org

Hypertext 2008

June 19-21, 2008 - Pittsburgh, Pennsylvania

JCDL 2008
June 16-20, 2008
Pittsburgh, PA

The Hypertext Conference provides the forum for research that considers links, their semantics, their presentation, the applications they have been put to, the knowledge that can be derived from their analysis and their effect on society. If information is connected, then the connection is called a link: the Hypertext Conference is concerned with all research concerning links.

Hypertext 2008 was an exciting and intellectually stimulating conference. The event featured more than 30 papers, posters, demos, and workshops which focused on key conference themes:

- Information Linking and Organization
- Social Linking
- Applications of Hypertext
- Hypertext, Culture, and Communication

Our keynote speakers examined leading-edge subjects in the linkage field. Fascinating presentations were offered by Bernardo Huberman of HP Labs and Prof. Jon Kleinberg of Cornell University. Paper topics ranged from assembling social semantic networks to dynamic modeling of tagging systems to enhancing access to open corpus educational content.

If you'd like to review the conference presentations, please visit the Conference Event links to the left. If you'd like to experience some of the fun at HT 2008, here are linking tools that will let you sample some of the presentations and view pictures from the conference.

Fig. 1 Hypertext 2008 Website: Original URI: http://ht2008.org, Current URI: http://www.sigweb.org/ht/ht08

results of our experiments on link neighborhood lexical signatures and it describes the parameters we tested for their generation. Section 7 provides aspects of future work and our conclusions of the article.

2 Related work

2.1 Inaccessible web resources

The Web is a highly dynamic environment with resources frequently changing over time. This fact has been subject to various studies over the years [12–18] and despite well-known guidelines for creating durable URIs [1], missing pages (HTTP response code 404) remain a pervasive part of the Web experience [19–23]. Over the last 15 years, numerous researchers have addressed the scale of the problem. A selection of related work quantifying the issue is given below, sorted by publication year.

1997: Kahle [24] found that the expected lifetime of a Web page was 44 days.

2000: Lawrence et al. [25] found that between 23 and 53 % of all URIs occurring in computer science related papers authored between 1994 and 1999 were invalid. By conducting a multi-level and partially manual search on the Internet, they were able to reduce the number of inaccessible URIs to 3 %. This confirms our intuition that information is rarely lost but rather moved to a different location.

2002: A study of Web page availability performed by Koehler [26] shows the random test collection of URIs eventually reached a "steady state" after approximately 67 % of the URIs were lost over a 4-year period. Koehler estimated that the half-life of a random Web page is approximately 2 years.

2003: Spinellis [27] conducted a similar study investigating the accessibility of URIs occurring in papers published in *Communications of the ACM* and *IEEE Computer Society*. He found that 28 % of all URIs were unavailable after 5 years and 41 % after seven years. He also found that in 60 % of the cases where URIs where not accessible, a 404 error was returned. He estimated the half-life of a URI in such a paper to be four years from the publication date.

Dellavalle et al. [28] examined Internet references in articles published in journals with a high impact factor (IF) given by the Institute for Scientific Information (ISI). They found that Internet references occur frequently (in 30 % of all articles) and are often inaccessible within 1 month after publication in the highest impact (top 1 %) scientific and medical journals. They discovered that the percentage of inactive references (references that return an error message) increased over time from 3.8 % after 3 months to 10 % after 15 months and up to 13 % after 27 months. The majority of inactive references they found were in the *.com* domain (46 %) and the fewest were in the *.org* domain (5 %). By manually browsing the IA, they were able to recover information for about 50 % of all inactive references.

2005: The work done by McCown et al. [29] focused on articles published in the *D-Lib Magazine*. Their results show a 10-year half-life of these articles. Nelson and Allen [30] studied object availability in digital libraries and found that 3 % of the URIs were unavailable after only 1 year.

2011: Sanderson et al. [31] studied the persistence and availability of Web resources that are referenced in scholarly articles. They found, for example, that approximately 25 % of all referenced URIs are not accessible anymore and are not available via Memento.

2.1.1 Soft 404s

Most of the previously mentioned studies determine the accessibility of a resource by testing the HTTP response code. If the request returns a 404 "Page not Found" response, it is obvious that the resource is inaccessible. However, Web servers occasionally respond to requests for inaccessible resources with the 200 response code (meaning "OK") along with a customized error page. This scenario is known as "soft 404" and Bar-Yossef et al. [32], for example, have proposed methods to identify them. Their idea was to send a second request to the suspected site with a string of random characters appended to the URI and to compare the content similarity of the two responses. Assuming that soft 404s responses are rather similar, regardless of the request, such sites could be isolated.

Lee et al. [33] took a more protocol-based approach by investigating the number and destinations of HTTP redirects for suspected soft 404 sites. Their assumptions is that, again regardless of the request, the final destination of (a chain of) redirects is identical for Web servers returning 200 instead of 404 responses.

Meneses et al. [34] showed that soft 404*s* can also be identified with text classifiers based on the characteristics of previously identified soft 404 pages. The authors were able to isolate lexical signatures of such pages, which contributed to predicting soft 404*s* with a precision of 99 % and a recall of 92 %.

2.2 Methods to overcome link rot

The scenario where links point to Web resources that have become unavailable is commonly referred to as link rot. The Hypertext Transfer Protocol (HTTP) [35] by default responds with the code 404 to a requested URI that cannot be found on the server. This response gives no indication of whether the erroneous condition is of a temporary or permanent nature. HTTP further provides the functionality to redirect a request to a page that has moved to a different location. The response code 301, for example, stands for a resource that has permanently been assigned a new URI. This response can result in an automatic redirect to the resource's new location. HTTP code 302, on the other hand, indicates a temporary move of a resource to a different URI. These procedures are helpful to avoid broken links if the Web administrator is aware of the actual new location of the page and modifies the configuration of the Web server accordingly. It is less useful for common Web users since they do not have this kind of administrative access to the server.

Several researchers have introduced methods to overcome the link rot problem that go beyond the native HTTP mechanisms. For example, Martinez-Romo and Araujo [36–38] have introduced a method to recover from link rot based mainly on querying the anchor text of links that point to the missing page against a search engine. They found that expanding the query with contextual data from the missing page (obtained from the Internet Archive) can improve the retrieval performance.

The work of Francisco-Revilla et al. [39] also addresses the issue of missing Web pages. They have developed a tool, which allows users to construct trails using Web pages which are usually authored by others. This path can be seen as a meta-document that organizes and adds contextual information to those pages. Thus, part of their research is about discovering relevant and significant changes to websites, with missing pages being a kind of change. Their evaluation of change is based on document signatures of paragraphs, headings, links and keywords. Just recently they redesigned the software ("Walden's Path") and launched version four of the system [40].

The work done by Harrison and Nelson [41] to find missing Web pages is closely related to our work as based on lexical signatures. They developed a server side system called OPAL which utilizes search engines to locate the desired page. OPAL kept a memory of references to avoid duplicate lookups but required administrator effort for Web server installation and configuration.

Popitsch and Haslhofer [42,43] introduced a tool called DSNotify to handle broken links in linked data environments. It was designed as a change detection framework as it monitors data environments, detects and attempts to correct broken links. It is capable of actively notifying subscribed applications of its actions. While DSNotify is well suited for bounded datasets, its applicability to Web scale environments remains unclear.

3 Lexical signatures

A textual document, for example a research paper, can consist of several thousands of words and hundreds of sentences. A paper usually also contains an abstract that summarizes the essence of the paper in no more than 300 words. An abstract, therefore, can be seen as the a reduced version of the paper. However, it is impractical to use the abstract or the full content as input for search engines. If the textual content of a document could be further reduced to, for example, less than ten terms, this reduction could be used as a search engine query. This small but meaningful representation of a textual document is a lexical signature. It can be compared to keywords as they are provided in research papers or meta tags in HTML documents. A lexical signature is light-weight metadata representing the content of a document.

3.1 Computation of lexical signatures

A lexical signature can be generated using the term frequency –inverse document frequency (TF–IDF) scheme [44]. This requires the extraction of all textual content of the document into a "bag of words". In related research [45], Web pages containing less than 50 words have been dismissed to ensure a good sized body of text to extract a lexical signature from. We adopted this filter for our experiments. Further, a language-dependent stop word filter was applied to dismiss all terms that do not contribute to the context of the document. We experimented with the application of stemming algorithms but dismissed them due to a drop in retrieval performance. The normalized TF value of term i is most simply computed following Eq. 1. This equation was introduced in [46,47] and it includes a smoothing factor a which is generally set to 0.4 [48].

$$\mathrm{TF}_{i_{\mathrm{norm}}} = a + (1-a)\frac{\mathrm{TF}_i}{\mathrm{TF}_{\max}} \tag{1}$$

However, various different normalization approaches have been introduced [49,50]. The IDF value of term i is computed following Eq. 2, where D denotes the total number of documents in the entire corpus and d_i is the number of documents in D that contain term i. Since it is possible that a term does not occur in any documents, which would lead to the division by zero, the denominator is frequently computed as $|d_i| + 1$. For our lexical signatures, we derived the values for $|d_i|$ from a search engine.

$$\mathrm{IDF}_i = log\frac{|D|}{|d_i + 1|} \tag{2}$$

The computation of IDF depends on global knowledge about the corpus, namely $|D|$ and $|d_i|$. If the entire Web is the corpus, these values cannot be computed accurately and have to be estimated. We have previously shown that using search engines for this estimation is a viable approach [51]. Values to estimate $|D|$ can be obtained from [52].

To compute a TF–IDF value for a term, its TF and IDF values are multiplied. The n terms with the highest TF–IDF value form our n-term lexical signature of the document.

3.2 Lexical signatures of web pages

Since the concept of lexical signatures of Web resources was first proposed by Phelps and Wilensky [53], little research has been done using lexical signatures for finding Web content that has moved from one URI to another. Phelps and Wilensky introduced the concept of "robust hyperlinks", a URI with a lexical signature of five terms appended as an argument. An example for a robust hyperlink is:

```
http://www.cs.berkeley.edu/~wilensky/
NLP.html?lexical-signature=
texttiling+wilensky+disambiguation+subt
opic+iago
```

where the lexical signature is the string following the "=" in the URI. This example was taken from Robert Wilensky's website. They conjectured that if the above URI would return a 404 error, the browser would take the appended lexical signature from the URI and automatically submit it to a search engine to find the page at its new location. Since Phelps and Wilensky's [54] goal was to find the same page, they set a TF value threshold and would not consider terms beyond that value for their lexical signatures. The lexical signature length of five terms was chosen somewhat arbitrarily.

Park et al. [45] expanded on the work of Phelps and Wilensky, studying the performance of nine different lexical signature generation algorithms (and retaining the 5-term precedent). They found that algorithms weighted for term frequency (TF) were better at finding related pages, but the exact page would not always be in the top n results. Algorithms weighted for inverse document frequency (IDF) were better at finding the exact page but were susceptible to small changes in the document (e.g., when a misspelling is fixed). The simple TF–IDF method described earlier that we used to generate our lexical signatures was one of the nine alternatives analyzed by Park et al. However, it is possible that a different lexical signature generation algorithm works better for applications different from ours.

3.3 Performance of n-term lexical signatures

Park et al. [45] evaluated several algorithms to generate lexical signatures but they did not investigate the relationship between the length and the age of a lexical signature and its retrieval performance. The purpose of our experiment was to address these questions left open by Park et al. and evaluate an optimal length for a lexical signature as a query for Web search engines and gain an insight into their decay over time.

We randomly sampled a relatively small set of 300 URIs from DMOZ. After applying an English language filter, dismissing pages with less than 50 words and dismissing pages without available Mementos, we were left with 98 URIs. We are aware of the limited size of this initial corpus but we expected it to grow significantly by obtaining all Mementos per URI for the experiment described in Sect. 3.5. In fact, the sample size grew to 10,493.

We generated ten lexical signatures of varying lengths for each of the 98 URIs and issued them against the Google search API. Since the Google API, at the time we conducted this experiment, had a limit of 1,000 queries per day, we only asked for the top 100 results. To evaluate the lexical signature performances, we parsed the result set of the individual queries and identified the URI the lexical signature was created from and its rank. The search results provided by the search engine APIs do not always match the result provided by the Web interfaces [55] but we used the Google API for all queries of this experiment and thus are not forced to handle possible inconsistencies. We distinguished between four retrieval scenarios for each URI. Either:

1. the URI is returned as the top ranked result or
2. the URI is returned in the top 10 but not as the top ranked result or
3. the URI is returned between rank 11 and 100 or
4. the URI is ranked somewhere beyond rank 100.

We considered a URI for the last case as undiscovered because numerous studies [56–60] have shown that the vast majority of Internet users do not look past the first few search results. These studies also show that users rarely click on search results beyond rank 10. We are aware of the potential discrimination of results ranked just beyond our threshold and there is an obvious difference between search results ranked 101 and, for example, rank 10,000. However, we chose this classification for simplicity and did not distinguish between ranks greater than 100.

Table 1 shows the performance statistics of all lexical signatures distinguished by their length in number of terms. It displays the relative amount of URIs returned in all of the four retrieval scenarios as well as the mean of these values (MR). The rightmost column holds the mean reciprocal rank (MRR) for the corresponding n-term lexical signatures.

Table 1 Lexical signature length vs. rank

	1	2–10	11–100	≥101	MR	MRR
2-Term	24.3	14.9	13.2	47.6	53.1	0.29
3-Term	40.2	15.0	**15.0**	29.8	36.5	0.45
4-Term	43.9	15.7	11.4	**29.0**	33.8	0.49
5-Term	47.0	**19.4**	3.4	30.2	**32.7**	0.52
6-Term	51.2	11.4	3.4	34.1	36.0	0.55
7-Term	**54.9**	9.4	1.5	34.2	35.5	**0.58**
8-Term	49.8	7.7	2.2	40.4	41.9	0.53
9-Term	47.0	6.6	0.9	45.5	46.4	0.50
10-Term	46.1	4.0	0.9	49.0	49.8	0.48
15-Term	39.8	0.8	0.6	58.9	59.5	0.40

The best values are in bold

The statistical significance (p value ≤ 0.05) of the results in Table 1 indicates three clusters. The first cluster is represented by 2-term lexical signatures, whose results are statistically significantly worse than all other lexical signatures. The second cluster contains the 3- to 8-term lexical signatures. Their results are similar but statistically better than the others. Finally, the third cluster contains the 9-, 10-, and 15-term lexical signatures. Their results are also similar to each other but worse than the second cluster and better than the first.

In general, we can observe a binary pattern meaning that the vast majority of URIs returned either ranked 1 or beyond 100. This pattern becomes even more obvious when comparing the top 10 results (including the top rank) and the number of undiscovered URIs. We see at most 15 % of URIs ranked between 11 and 100.

The first result of this experiment is that 7-term lexical signatures performed best. They showed the best results in terms of most top ranked URIs as well as in terms of MRR. We consider this a refinement of the assumptions made by Phelps and Wilensky and Park et al.

5-Term lexical signatures returned fewer URIs top ranked and their MRR was lower as well. They did, however, show the best mean rank and returned the most URIs in the top 10. Both facts support the initial preference and show that this lexical signature length can return very good results. The performance of 6-term lexical signatures falls somewhere between 5- and 7 terms. Even though they returned more top ranked URIs compared to 5-term lexical signatures, they also left more URIs undiscovered. The performance of 2-term lexical signatures was rather poor and 3- and 4-term lexical signatures also are no competition to 5 or 7 terms. The picture for 8-, 9- and 10-term lexical signatures is basically the same. Their performance was not very impressive and got worse as more terms were added as the values for 15-term lexical signatures prove. The binary pattern, however, is best visible at these high-term lexical signatures.

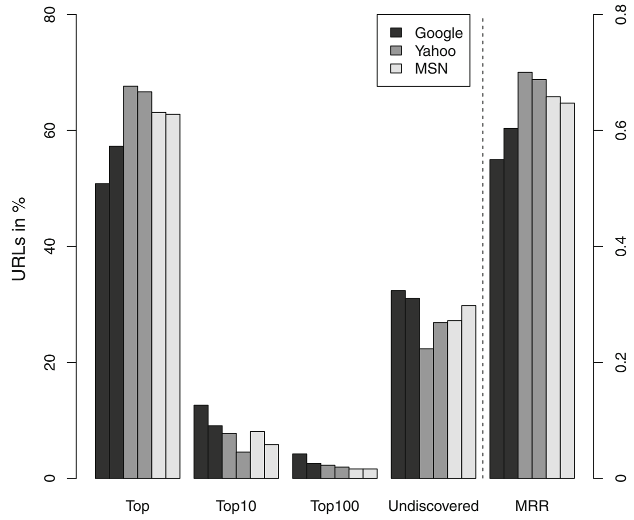


Fig. 2 5- and 7-Term lexical signature retrieval performance

3.4 Performance between search engines

In the previous experiment, the IDF values were derived from a single search engine, Google. In this follow-up experiment, we compared the performance of 5- and 7-term lexical signatures generated based on $|d_i|$ values obtained from three different search engines. We used the Google, Yahoo! (BOSS) and MSN Live APIs to determine IDF values and estimated $|D|$ with the help of [52]. We obtained a larger data set by randomly sampling 500 URIs from the Open Directory Project. After applying the common filters described above and in [45], our final sample set consisted of a total of 309 URIs, 236 in the .com, 38 .org, 27 .net and 8 in the .edu domain. For each URI, we computed a 5-term and a 7-term lexical signature per search engine (with the same simple TF–IDF method described earlier) meaning we created a total of six lexical signatures per URI. We queried each lexical signature against the search engine it was based on and applied the previously introduced four retrieval scenarios.

Figure 2 shows the percentage of URIs retrieved top ranked, ranked in the top 10, in the top 100 as well as the percentage of URIs that remained undiscovered when using 5- and 7-term lexical signatures. For each of the four scenarios, three tuples are shown distinguished by color, indicating the search engine the lexical signature was generated from and queried against. The left bar of each tuple represents the results for 5- and the right for 7-term lexical signatures. The rightmost set of columns represents the MRR of the corresponding lexical signature lengths and it refers to the right *y* axis which shows a normalized scale.

The best performance is observed for the 5-term lexical signature derived from Yahoo!. It retrieves 67.6 % of all URIs in our sample set top ranked, 7.7 % ranked in the top 10 (but not top) and 22 % remain undiscovered. We can observe the binary pattern again with the majority of the URIs either returned in the top 10 (including the top rank) or remaining undiscovered, across search engines and query lengths. More than 75 % of all URIs are ranked between one and ten and the vast majority of the remaining quarter of URIs was not discovered.

Yahoo! returned the most URIs and left the least undiscovered. MSN Live, using 5-term lexical signatures, returned more than 63 % of the URIs as the top result and hence performed better than Google which barely returned 51 %. Google was the only search engine returning more top ranked results with 7-term lexical signatures and it showed more URIs ranked in the top 10 and top 100 compared to Yahoo! and MSN.

These results suggest the use of the Yahoo! BOSS API for our further experiments.

3.4.1 Cross-search engine performance

The previous results raise the question about the dependency between IDF values derived from one search engine when used to query another. To investigate this relationship, we took all previously generated 5-term lexical signatures and queried them against all three search engines and not just the index they were generated with.

Figure 3 shows the 5-term lexical signature performance in all three search engines. The labels on the axes indicate what search engine the lexical signatures were derived from (first letter) as well as what search engine they were queried against (second letter). G, M and Y stand for Google, MSN and Yahoo! respectively. The label GM, for example, represents lexical signatures based on Google and queried against MSN. The size of the circles is proportional to the number of URIs returned. The absolute values are also plotted in the graph, either inside or right next to the corresponding circle. We again distinguish between our four retrieval scenarios.

Lexical signatures derived from Yahoo! performed best when queried against Yahoo! Even though YG returned almost twice as many URIs in the top 10 than YY, its performance in the top ranks was much worse.

Lexical signatures derived from Google left the least URIs undiscovered when queried against Google. However, the performance in the top ranks was better when queried against Yahoo! and MSN.

Lexical signatures derived from MSN performed better when queried against Yahoo! (MY) or Google (MG) than against MSN itself (MM). They returned more top ranked URIs or URIs in the top 10 and top 100 and left fewer URIs undiscovered.

These results indicate that, especially when utilizing the Yahoo! Boss API, querying the lexical signatures against the very same index is preferable. This is the third result of the lexical signature experiment.

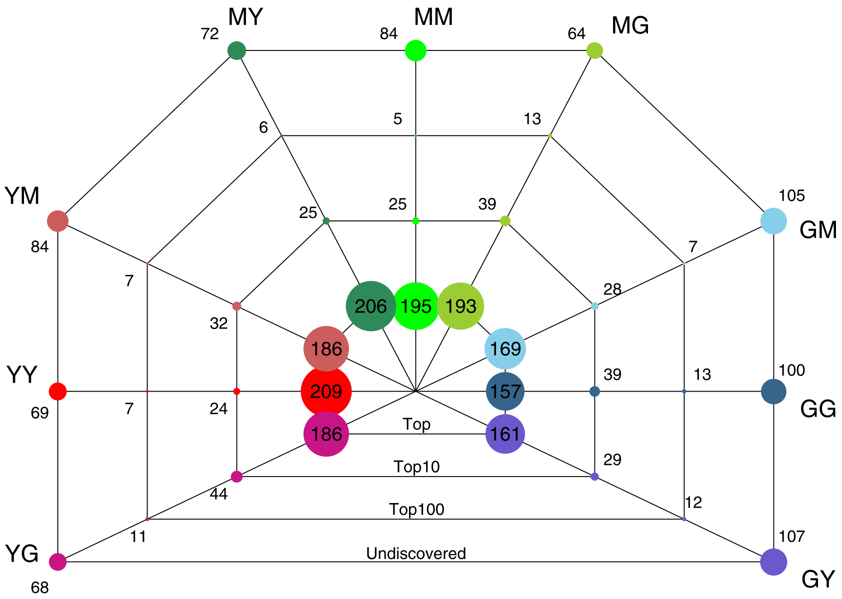


Fig. 3 Lexical signature performance across three search engines

3.5 Evolution of lexical signatures over time

In Sect. 2, we have provided numerous references to related research showing that Web content changes frequently over time. Consequently, Web page lexical signatures are prone to change as well. In this section, we provide some insight into lexical signature evolution over time and the effect on their retrieval performance.

Table 2 shows various example lexical signatures created at different points in time. The first three lexical signatures were created by Phelps and Wilensky in the late 1990s. The two lexical signatures for the Endeavour project at Berkeley from the 1990s and 2011 share two out of five terms. Interestingly, the zip code has made it into the recently created lexical signature even though the content of the page has not changed in the last 11 years. This most likely is due to the increased size of the corpus (the Web) where a nine-digit number became a better discriminator against other more common terms. The lexical signatures for Randy Katz's homepage in contrast do not show any term overlap. Correcting the typo in the word *California* on the Web page likely contributed to the disappearance of the term from the lexical signature since *California* is not a good discriminator in the entire index of a modern search engine. The lexical signature of the Web page for the Digital Libraries Initiative 2 was also created in the late 1990s. Today, the URI returns a 404 error – the project has expired years ago. The recent lexical signature was created from the last available Memento provided by the IA from 2009. We see no overlap between the two lexical signatures. The lexical signatures of the Library of Congress example have three terms in common, all of which one would expect to find on this website. The JCDL 2008 example shows the highest overlap with four terms. Only the email address replaced the less discriminating token *pst*.

Table 3 shows the results of querying the lexical signatures of the URIs shown in Table 2 at different points in time. The query results of the lexical signatures by Phelps and Wilensky in the late 1990s can be obtained from their numerous presentations available on the Web.

We can see that all lexical signatures created in the past performed very well in the past. All three lexical signatures by Phelps and Wilensky showed an excellent performance by returning the URI top ranked. Two of the three lexical signatures return the target URI as the only result. The lexical signature of the Library of Congress had a high recall but still returned the URI top ranked. Even though the JCDL URI was only returned ranked second, given the low recall value (only 77 total results), the lexical signature can still be considered well performing.

Querying the old lexical signatures today shows a different picture. We did not find the DLI2 URI or the URI of Randy Katz's page. The DLI2 URI no longer exists and since it has been deleted from the search engine's index it could not be returned. The URI of Randy Katz's page, however, is still indexed and could have been returned. The three URIs that were returned were ranked in the top 10 which is a good result even though the recall is rather high. The newly generated lexical signatures performed much better with all indexed URIs returned top ranked and a low recall value with the Library of Congress lexical signature being the exception.

These examples show that the performance of lexical signatures changes over time. An up-to-date lexical signature performs better in the sense of finding recent versions of a page. However, an old lexical signature could still be used for identifying an old version of the page.

Figure 4 displays the normalized Discounted Cumulative Gain (nDCG) [61] values of select lexical signatures we created over time using the Memento framework [3]. Each data

Table 2 Lexical signatures generated from various URIs over time

URI	Lexical signature	
	Past	Recent
http://endeavour.cs.berkeley.edu/	Amplifies endeavour leverages charting expedition (late '90s)	Endeavour 94720-1776 achieve inter-endeavour amplifies (2011)
http://bnrg.eecs.berkeley.edu/~randy/	californa isrg culler rimmed gaunt (late '90s)	randy eecs professor frameset katz (2011)
http://www.dli2.nsf.gov/	nsdl multiagency imls testbeds extramural (late '90s)	digital library dli2 2002 2003 (2009)
http://www.loc.gov/	library collections congress thomas american (2008)	library librarian congress webcasts collections (2012)
http://www.jcdl2008.org/	libraries jcdl digital conference pst (2008)	libraries jcdl digital conference info@jcdl2008.org (2011)

Table 3 Lexical signatures generated from URIs over time queried against Google at different points in time. Results are shown as rank/total results (year of the query)

URI	Past LS Queried		Recent LS
	in Past	Recently	Queried Recently
http://endeavour.cs.berkeley.edu/	1/1 (late '90s)	4/194,000 (2011)	1/139 (2011)
http://bnrg.eecs.berkeley.edu/~randy/	1/<100 (late '90s)	NA/11 (2011)	1/9,340 (2011)
http://www.dli2.nsf.gov	1/1 (late '90s)	NA/19 (2011)	NA/8,670 (2011)
http://www.loc.gov	1/174,000 (2008)	2/356,000 (2011)	1/762,000 (2012)
http://www.jcdl2008.org	2/77 (2008)	9/550 (2011)	1/617 (2011)

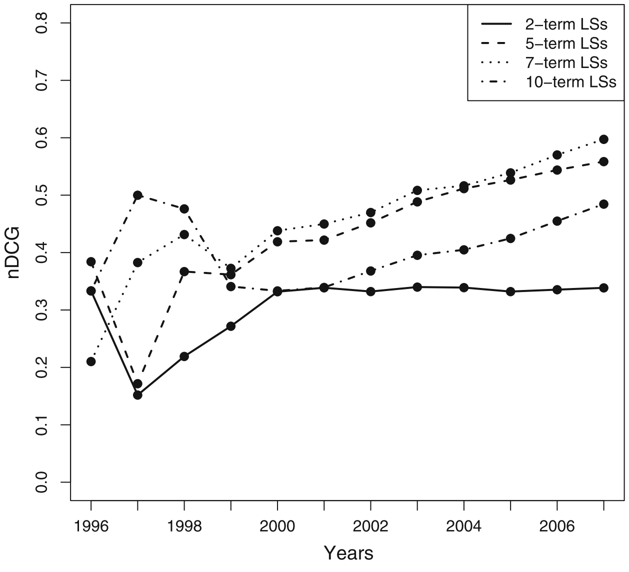


Fig. 4 Lexical signature performance over time

point represents the mean nDCG score of all URIs of a certain year indicated by the values on the x-axis. The great fluctuation of the numbers for the early years in Fig. 4 can be explained with the limited number of Mementos per URI for that time. We do believe, however, that from roughly year 2,000 on there is a pattern visible.

Figure 4 confirms the top performance of 5- and 7-term lexical signatures but also shows that lexical signatures older than four to five years perform poorly. This is the fourth result of our series of lexical signature experiments. For a more in-depth study of the evolution of lexical signatures, we refer to the author's dissertation work [4].

3.6 Results

The series of experiments on the performance of lexical signatures has four main results:

1. With respect to the preferred length of a lexical signature, we have shown that our 7-term lexical signatures outperformed their 5- and 6-term counterparts. The performance and applicability of this method are dependent on the availability of Mementos of missing pages because without them, no lexical signature can be generated.
2. We have seen indicators that suggest the use of the Yahoo! BOSS API for our further experiments. The API showed the best performance to derive IDF values and to query the generated lexical signatures against compared to Google and MSN.
3. Related to that, we have shown indicators that, when having a lexical signature derived from the Yahoo! Boss API, querying it against the very same index is preferable. The

Table 4 Example of well-performing lexical signatures and titles obtained from two different URIs

		Rank
URI	http://www.aircharter-international.com	
LS	*Charter Aircraft Jet Air Evacuation Medical Medivac*	1
Title	*ACMI, Private Jet Charter, Private Jet Lease, Charter Flight Service: Air Charter International*	1
URI	http://www.nicnichols.com	
LS	*NicNichols Nichols Nic Stuff Shoot Command Penitentiary*	1
Title	*NicNichols.com: Documentary Toy Camera Photography of Nic Nichols: Holgs, Lomo and Other Lo-Fi Cameras!*	1

performance decreases when querying the lexical signature against a different search engine.

4. With respect to the evolution of lexical signatures over time and the impact on its retrieval performance, we found that the performance of our top lexical signatures (5- and 7-terms of length) drops dramatically if they are older than 4–5 years. This means that chances to rediscover a missing page based on a lexical signature generated from a recent Mementos are higher than if it was derived from a 5 year old Memento.

4 Titles

We have seen that lexical signatures can perform well for discovering missing Web pages. However, their generation, following the TF–IDF scheme, is expensive. In this section, we describe our experiments on the performance of Web page titles as a cheaper method to obtain a search engine query. We also analyzed the gain when combining the title and the lexical signature methods. We further investigated the evolution of titles over time and compared it to the evolution of document content over time. We maintained a few underlying assumptions regarding Web page titles. We anticipated that a majority of Web pages actually have titles and believed that the titles are descriptive of page content.

To illustrate the concept behind this experiment, we show two examples in Table 4. It displays the titles and lexical signatures obtained from two URIs. When queried against Google, both the titles and the lexical signatures return the corresponding URI top ranked. This example is promising and motivated us to further investigate the retrieval performance of Web page titles.

4.1 Title extraction

Researchers such as Chakrabarti et al. [62] have found (in a corpus of 1 million URIs) that up to 17 % of HTML documents lack titles. While this is a high percentage, it leaves more than 80 % of Web pages with titles which for us justifies further investigating this method. In a brief and somewhat brute force experiment, we randomly picked 10, 000 URIs from DMOZ and found that only 1.1 % of URIs lack a title. This confirms our intuition that titles of Web pages are commonplace. However, it also confirms the potential bias of sampling from DMOZ as discussed in Sect. 1. The URIs are curated and, therefore, less likely to be missing distinguishing features such as titles We used the same sample set of 309 URIs introduced in Sect. 3.4 and obtained the titles of all Web pages by extracting the content of the HTML element $< title >$.

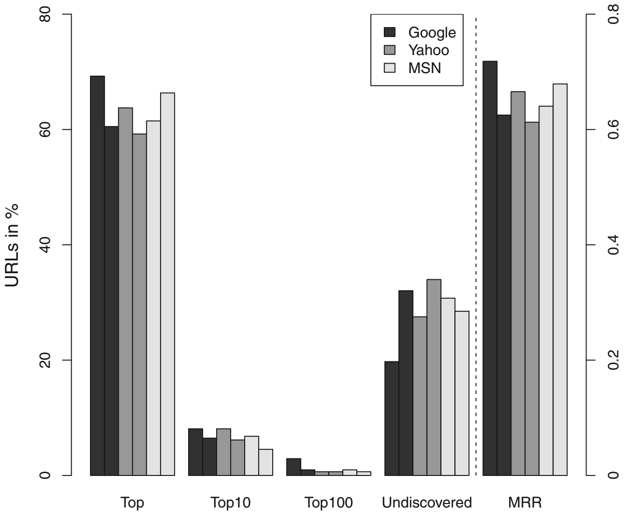


Fig. 5 Non-quoted and quoted title retrieval performance

4.2 Performance of Titles

Similar to the experiment described in Sect. 3.4, we issued queries against Google, MSN Bing and Yahoo!. We also evaluated the results by distinguishing between our four retrieval scenarios.

Figure 5 shows the percentages of retrieved URIs when querying the title of the page. We queried the title once without quotes and once quoted, forcing the search engines to handle all terms of the query as one string. Each tuple is distinguished by color and the left bar shows the results for the non-quoted titles. The rightmost set of columns represents

Table 5 Examples for well and poorly performing lexical signatures and titles

		Rank
URI	http://www.redcrossla.org	
LS	*Marek Halloween Ready Images Schwarzenegger Governor Villaraigosa*	>100
Title	*American Red Cross of Greater Los Angeles*	1
URI	smiledesigners.org	
LS	*Dental Imagined Pleasant Boost Talent Proud Ways*	1
Title	*Home*	>100

the MRR of the corresponding titles and it refers to the right y axis which shows a normalized scale.

Figure 5 reveals the top performance of titles when queried non-quoted against Google with 69.3 % URIs top ranked. It is surprising to see that both Google and Yahoo! returned fewer URIs when using quoted titles. Google, in particular, returned 14 % more top ranked URIs and 38 % fewer undiscovered URIs for the non-quoted titles compared to the quoted titles. Only MSN Live showed a different behavior with more top ranked results (almost 8 % more) for the quoted and more undiscovered URIs (more than 7 %) using the non-quoted titles. Figure 5 represents the first result of this experiment based on our sampled URIs: titles are a very well-performing alternative to lexical signatures. Recall that the top value for lexical signatures taken from Fig. 2 was obtained from Yahoo! (5-term) with 67.6 % top ranked URIs returned.

4.3 Combined title and lexical signature performance

Titles are usually created by humans which intuitively makes us understand that not all titles are equally good. The examples displayed in Table 5 illustrate the potential differences between the retrieval performance of titles and lexical signatures. In this section, we describe our experiment to investigate the possible gain from combining both methods.

The first example in Table 5 shows the lexical signature and the title obtained from the URI http://www.redcrossla.org. The lexical signature represents the content of the page at a certain point in time rather than describing the general “aboutness”. Hence the page was not returned in the result set of a Google search. The title of the page, however, captures the timeless essence of the Web page of the Red Cross in Los Angeles and consequently performed much better and returned the URI top ranked. This example illustrates that despite the reliable TF–IDF based selection of the most salient terms of a page, a lexical signature is not automatically the best chosen query string. A Web page’s title can be more robust since a title is understood to capture the overall topic of a page or a document. The second example represents data taken from the URI smiledesigners.org, a Web page of a dentist. The generated lexical signature returned the URI top ranked. However, the title is an unfortunate choice. While *Home* may be a good title within the site, it does not distinguish this page from many others on the Web. Submitted to Google, it did not return the URI within the top 100 results (but it was indexed with the term). This example shows that not all titles are equally good for Web retrieval. Results of our detailed study on the quality of Web page titles can be found in [63].

To analyze the potential gain from combining both methods, we modified the previous experiment. We defined three queries per URI: its title, its 5-term, and its 7-term lexical signature. The lexical signatures were computed based on the same TF–IDF method detailed in the previous section and the $|d_i|$ values were derived from the Yahoo! BOSS API. The methods were combined in a way where the first method is applied to all URIs. For those URIs that remained undiscovered, a second method was applied and for URIs that still remained undiscovered, the third method was applied. This implies that the order of methods matters. Table 6 shows all reasonable combinations of all three queries. $LS5$ and $LS7$ stand for 5- and 7-term lexical signatures and TI stands for title queries. The top performing methods are highlighted in bold figures (one per row).

Regardless of the sequence of methods, the best results were obtained from Yahoo!. If we consider all combinations of only two methods, we find the top performance of 75.7 % twice in the Yahoo! results. Once with $LS7 - TI$ and once with $TI - LS5$. The second result of the title experiment is the recommendation for the use of the $TI - LS5$ sequence. This point is mainly supported by two reasons:

1. titles are easier to obtain than lexical signatures, and
2. this methods returned 9.1 % of the URIs in the top 10 which is 1.7 % more than the sequence $LS7-TI$ returns. Even though we do not distinguish between rank two and rank nine, we still consider URIs returned within the top 10 as good results.

The sequence $LS7 - TI - LS5$ accounts for the most top ranked URIs overall with 76.4 %. While the 3-method sequence returned good results, they were not drastically better than, for example, the two methods mentioned above. The performance delta was not sufficient to justify the expensive generation of lexical signatures without using the easy to obtain titles first.

Table 6 Relative number of URIs retrieved with two or more methods combined

	Google				Yahoo!				MSN Live			
	1	10	100	>100	1	10	100	>100	1	10	100	>100
LS5-TI	65.0	15.2	6.1	13.6	**73.8**	10.0	2.3	14.0	71.5	10.0	1.9	16.5
LS7-TI	70.9	11.7	4.2	13.3	**75.7**	7.4	1.9	14.9	73.8	9.1	1.9	15.2
TI-LS5	73.5	9.1	3.9	13.6	**75.7**	9.1	1.3	13.9	73.1	9.1	1.3	16.5
TI-LS7	74.1	9.4	3.2	13.3	**75.1**	8.7	1.3	14.9	74.1	9.1	1.6	15.2
LS5-TI-LS7	65.4	15.2	6.5	12.9	**73.8**	10.0	2.6	13.6	72.5	10.4	2.6	14.6
LS7-TI-LS5	71.2	11.7	4.2	12.9	**76.4**	7.8	2.3	13.6	74.4	9.1	1.9	14.6
TI-LS5-LS7	73.8	9.1	4.2	12.9	**75.7**	9.1	1.6	13.6	74.1	9.4	1.9	14.6
TI-LS7-LS5	74.4	9.4	3.2	12.9	**75.7**	9.1	1.6	13.6	74.8	9.1	1.6	14.6
LS5-LS7	52.8	12.9	6.5	27.8	**68.0**	7.8	2.9	21.4	64.4	8.4	2.6	24.6
LS7-LS5	59.9	9.7	2.6	27.8	**71.5**	4.9	2.3	21.4	66.7	7.1	1.6	24.6

The best values are in bold

Yahoo! uniformly gave the best results and MSN Live was a close second. Google was third, only managing to outperform MSN Live once ($TI - LS5$) at the top rank.

4.4 Title evolution versus document change

It is our intuition that Web page titles change less frequently and less significantly than Web page content. The title supposedly reflects the general topic of a page, which naturally changes less often than its content. If this intuition is correct, a title could constitute a reliable and easy to obtain search engine query for discovering missing Web pages.

To assess this intuition, we conducted an experiment based on a new and much larger data set. We randomly sampled 20,000 Web pages from DMOZ and after applying the same filters as described in the previous section, we were left with almost 7,000 pages. To investigate the evolution of titles over time, we queried each URI against the IA for Mementos [3] (old copies). For a total of 6, 093 URIs from DMOZ, we obtained a TimeMap (a list of Mementos for an original URI). We downloaded all available Mementos from 1996 until 2011 (more than 500,000) and extracted the page content and title.

To assess the level of content similarity between Mementos for a URI, we computed shingle values for all of them. We normalized these values so that zero indicates a very similar page and one represents very dissimilar page content. We then took the average over all Mementos per URI. We used the Levenshtein [64] edit distance for a similarity measure between all titles of all Mementos. The Levenshtein edit distance conveys how many operations are needed to transform on string into another and hence it is very suitable for title strings. We also took the average of the Levenshtein edit distance over all Mementos per URI.

Figure 6 shows the average normalized edit distance on the x axis and the average normalized shingle value of the same URI on the y axis. Both values are rounded to the nearest tenth. The color indicates the amount of times a certain point was plotted at the same coordinates. The palette starts with a basic green indicating a frequency of less or equal than 10 and transitions into a solid red representing a frequency of more than 90. The semi-transparent numbers represent the total amount of points in the corresponding quarters and their halves. The pattern is very apparent. The vast majority of points were plotted with an average shingle value of above 0.5 and an average edit distance of below 0.5. That translates to a high title similarity and a high content dissimilarity at the same time for the majority of the URIs. In fact, the most frequently plotted point was plotted more than 1, 600 times. It is (as an exception) colored black and located at the coordinates [0, 1] meaning close to identical titles and very dissimilar content. The point at [0, 0] was plotted 122 times and hence somewhat significant as much as some points with a shingle value of one and an edit distance of above 0.5. These points have transitioned to red.

Figure 6 supports our intuition that titles change less significantly over time than page content—our third result for the experiment. Given the dominant frequency of the point that represents identical titles and very dissimilar content, we are led to believe that titles, compared to lexical signatures, are the more robust retrieval method for discovering missing Web pages.

4.5 Results

The series of experiments on the retrieval performance of titles has three main results:

1. Titles are a very well-performing alternative to lexical signatures as their retrieval performance is very similar to lexical signatures (shown in Sect. 3). In addition, they are easy to obtain (by extracting the content of the HTML element <title>) and do not, unlike lexical signatures,

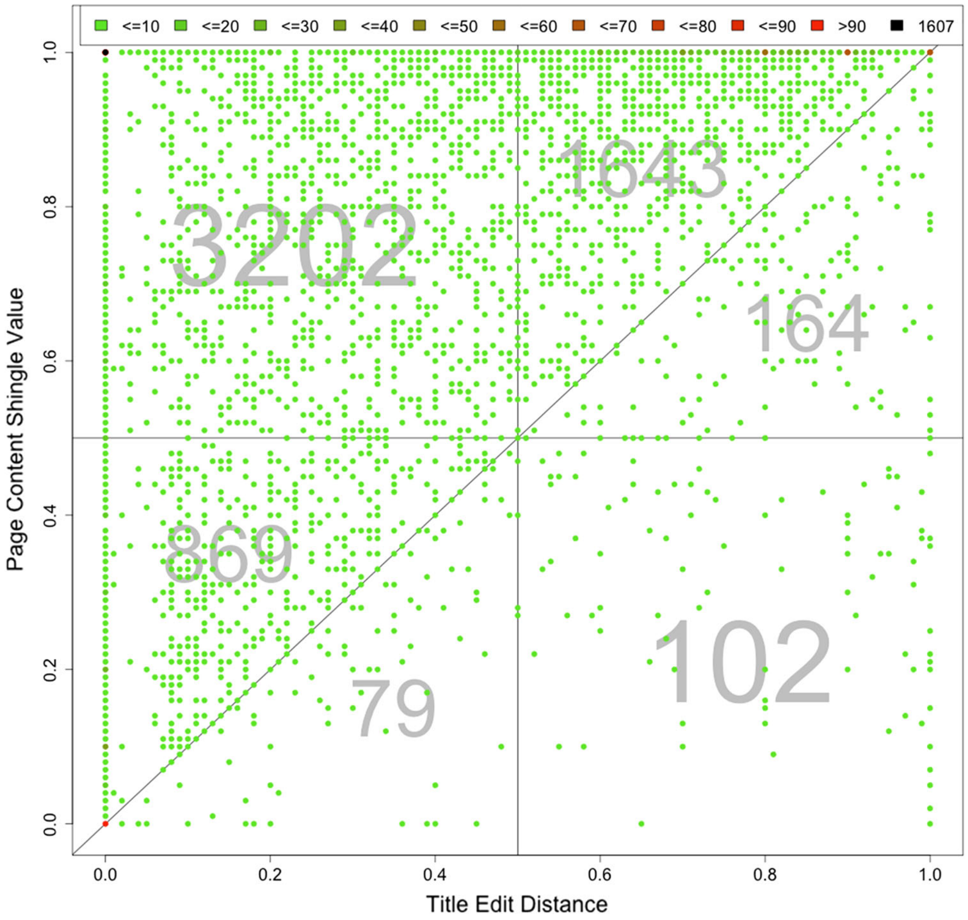


Fig. 6 Title edit distance and document changes of URIs

require the computation of TF–IDF values. The performance and applicability of this method are dependent on the availability of Mementos of missing pages because without them, no title can be extracted.

2. Combining methods can improve retrieval performance. The sequence $TI - LS5$ performs best and given that titles are cheap to obtain, they should be applied as the first method. The combination of methods is also only feasible if Mementos are available.
3. With respect to the evolution of titles over time, we have shown evidence that titles change less significantly over time than Web page content. This means that even in a case where only old Mementos of a missing page are available, chances to rediscover the page using its title are better than using its lexical signature.

5 Tags

The third method we investigated for rediscovering missing Web pages is the use of tags. Tags, as a form of user-generated metadata about Web pages, have been shown to be suitable for Web search. For example, Bao et al. [65] have observed that tags from the social bookmarking site *Delicious* are usually good summaries of the corresponding Web pages. Jason Morrison [66] also investigated the usefulness of tags for search and found in an extensive study that search in folksonomies can be as precise as search in major modern Web search engines. These results are confirmed by Heymann et al. [67], who found that tags significantly overlap with popular search terms, indicating that tags can indeed help locating relevant pages. Another intriguing result was shown in the work by Bischoff et al. [68]. According to their results, more than 50 % of tags annotating an URI do not occur in the content of the corresponding Web pages. That implies that tags provide additional information, which in fact can be useful for Web search.

Unlike the two previously introduced methods, the use of tags is applicable even if no Mementos [3] of a missing Web page exist. Tags, hosted by various different services in the Web, may very well outlive the page they annotate.

5.1 Performance of tags

We analyzed our existing corpora and found that URIs with tags were very sparse. We only found tags for about 15 % of

Table 7 Relative retrieval numbers for tag-based query lengths in number of tags

# of Tags	Top	Top10	Top100	Undis	MR	MRR
4	7.2	11.3	9.6	71.9	76.3	0.12
5	9.0	11.3	**9.7**	69.7	74.2	0.13
6	9.7	**12.0**	9.0	**69.3**	73.4	0.14
7	10.5	11.5	8.7	**69.3**	**73.1**	**0.15**
8	**11.0**	10.8	8.1	70.1	73.6	**0.15**
9	10.3	9.9	8.0	71.9	75.2	0.14
10	9.7	8.9	6.4	75.0	78.0	0.13

The best values are in bold

all URIs and other researchers such as Heymann et al. [67] made the same observation. Given this observation, we did not expect tags to outperform titles and lexical signatures. We assumed, however, that tags, if available, combined with titles and lexical signatures could provide an added value for rediscovering missing Web pages.

To generate a meaningful corpus to research tags we generated a new, "tag-centric" corpus. At the time, this experiment was conducted, the website delicious.com provided the best source for obtaining tags and the URIs they annotate. Since then the operation of Delicious has changes and hence obtaining their tags is not as easy anymore as it was at the time this experiment was run. We aggregated 4, 968 unique URIs from the Delicious index using their "random tool".[2] We are aware of the bias of our dataset towards the Yahoo! index (which we queried against), especially in the light of Yahoo! integrating Delicious data into their index [69]. However, sampling from Delicious was a popular approach taken by various researchers [68,67]. We used screen scraping, instead of the Delicious API, to gather up to 30 tags per URI. As previously shown [70], the Delicious API is unreliable, which was the main reason for this decision. The order of Delicious tags, which may be of relevance for Web search, indicates the frequency of use for all tags.

We first analyzed the retrieval performance of tag-based queries in terms of the number of tags they contain. Table 7 shows query lengths varying from 4 to 10 tags and their performance in relative numbers with respect to our four retrieval categories plus the mean rank and MRR. It shows that 8-tag queries returned the most top ranked results (11 %) and 7-tag queries, tied with 6-tag queries, left the fewest URIs undiscovered. However, results from all tag-based queries shown in Table 7 are very similar, regardless of the query length in number of tags. In fact, we could not find a statistical significance (p value ≤ 0.05) between any of the results and hence we cannot confidently promote one query length over another. These two observations form the first result of this experiment.

5.2 Combining tags with other methods

Table 7 shows that the overall retrieval performance of tags alone was not impressive. This lead us to investigate how the union of the results of more than one method would improve the retrieval performance.

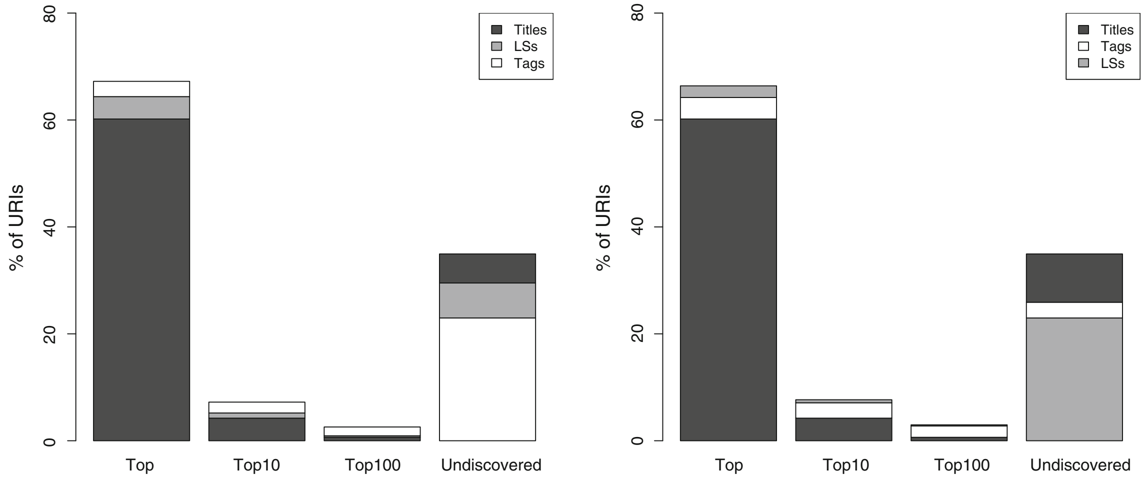
Extracting a Web page's title from the content is cheap; it costs just one request to the resource. In case the resource is a Memento it entails (in the simplest case) two requests: one to locate the Memento and the second to obtain the archived resource itself. Lexical signatures are much more expensive to generate. An TF–IDF value needs to be computed for each term, which entails one request per unique term plus the computation of TF values. Obtaining tags, similar to titles, is very cheap because it only requires one request per URI.

With this "cost model" in mind, we defined two sequences of methods to form our queries: *Title-Lexical_Signature-Tags* (*T-LS-TA*) and *Title-Tags-Lexical_Signature* (*T-TA-LS*). Since titles performed best (as shown in Sect. 4 and also demonstrated in previous work [71]), we maintained the priority for titles and queried them as our first step in both sequences. As the second step in *T-LS-TA*, we applied the lexical signature based method to all URIs that remained undiscovered. The third step was to apply the tag-based method to all URIs that were still undiscovered. The difference in the second sequence was that the tag-based method was applied second and the lexical signature based method third.

Figure 7 shows the combined retrieval performance of both sequences. The data of sequence *T-LS-TA* are shown in Fig. 7a. The previously introduced four retrieval categories are shown and the contribution per method is distinguished by grey scale. The first three bars (from left to right) are additive, meaning that the darkest part of the bars corresponds to the relative number of URIs returned by titles, the gray portion of the bars corresponds to the URIs not returned by titles but returned by lexical signatures. The white part of the bars represents the URIs neither returned by titles nor by lexical signatures but by tags only. Therefore, these three left bars are to be read as if they were growing with the application of each additional method. The rightmost bar is to be read as if it was subtractive. For Figure 7(a), it means the dark portion of the bar represents the number of URIs undiscovered with titles (34.9 %). The upper bound of the dark portion down to the upper bound of the gray portion represents the retrieval gain achieved by applying the second method. The height of the white portion of the bar corresponds to the final number of URIs that were left undiscovered after applying all three methods (23%) in the sequence *T-LS-TA*.

Figure 7(b) displays the data in the same way for the sequence *T-TA-LS*. The scheme of the grey scale remains

[2] http://www.delicious.com/recent/?random=1.



(a) Title-Lexical Signature-Tag Sequence

(b) Title-Tag-Lexical Signatures Sequence

Fig. 7 Performance of titles combined with lexical signatures and tags

the same with respect to the method meaning dark is still the title, gray still the lexical signature and white still represents tags. The height of the gray bar for undiscovered URIs is identical to the corresponding white bar in Fig. 7a. The additive bar for the top ranked results is slightly higher in Fig. 7a (67.2 vs. 66.4 %) but the bars for the top 10 and top 100 results are slightly higher in Fig. 7b (7.2 vs. 7.7 % and 2.6 vs. 3.0 %).

These results show that adding tags to the sequence of retrieval methods can improve the overall results. As long as tags are available, they performed similarly to lexical signatures as a secondary method. Since tags are much cheaper to obtain, if possible, we recommend the *T-TA-LS* sequence for rediscovering missing Web pages. This is the second result of the experiment.

5.3 Ghost tags

Previous research [68,67] has shown that about half the tags used to annotate URIs do not occur in the page's content. We found a slightly higher value with 66.3 % of all tags not present in the pages of our Delicious-based corpus. However, these numbers only apply for the current version of the page. The tags provided by Delicious on the other hand were aggregated over an unknown period of time (at the time the experiment was conducted, tags in Delicious could not be accurately dated). This means that it is possible that some tags used to occur in the content of a previous version of a page (a Memento) but were removed from it at some later point. However, through Delicious the tags of that page are still available. We call these tags “ghost tags” as they are terms that persist as tags after disappearing from the document itself.

To further investigate this aspect, we used the Memento framework [3] to obtain TimeMaps for all URIs that have tags not occurring in their content. For our dataset, this applied to more than 95 % of the URIs. Since we obtained different amounts of Mementos and different ages of the Mementos, we decided to only check tags against the first Memento meaning the oldest available copy of the page. We obtained TimeMaps for 3, 306 URIs, some of which date back to 1996. Out of all tags not present in the current page (66.3 % of all tags) we found a total of 4.9 % being ghost tags, meaning that they appeared in the first Memento.

These observations confirm that ghost tags exist, meaning that some tags better represent the past content of a Web page than the current. They do not, however, give indicators about the importance of ghost tags for the document and for the user. To further analyze this aspect, we compared the tags' frequency of use-based rank in Delicious (as a measure of importance to the user) with its TF-based rank in the first Memento (as a measure of importance to the document at the time). We normalized the ranks to a value between zero and one to avoid a bias towards a greater amount of available tags and longer documents. The closer the value gets to zero the higher is the rank, meaning the greater the tag's importance.

Figure 8 displays the Delicious rank on the x axis and the TF rank on the y axis. Each dot represents one ghost tag. If a dot is plotted more than once, its shade gets darker. 18 dots are plotted twice, one is plotted three times and one five times. The semi-transparent numbers indicate the percentage of ghost tags in the corresponding quadrants. The numbers show a majority of ghost tags (34.7 %) occurring in the first

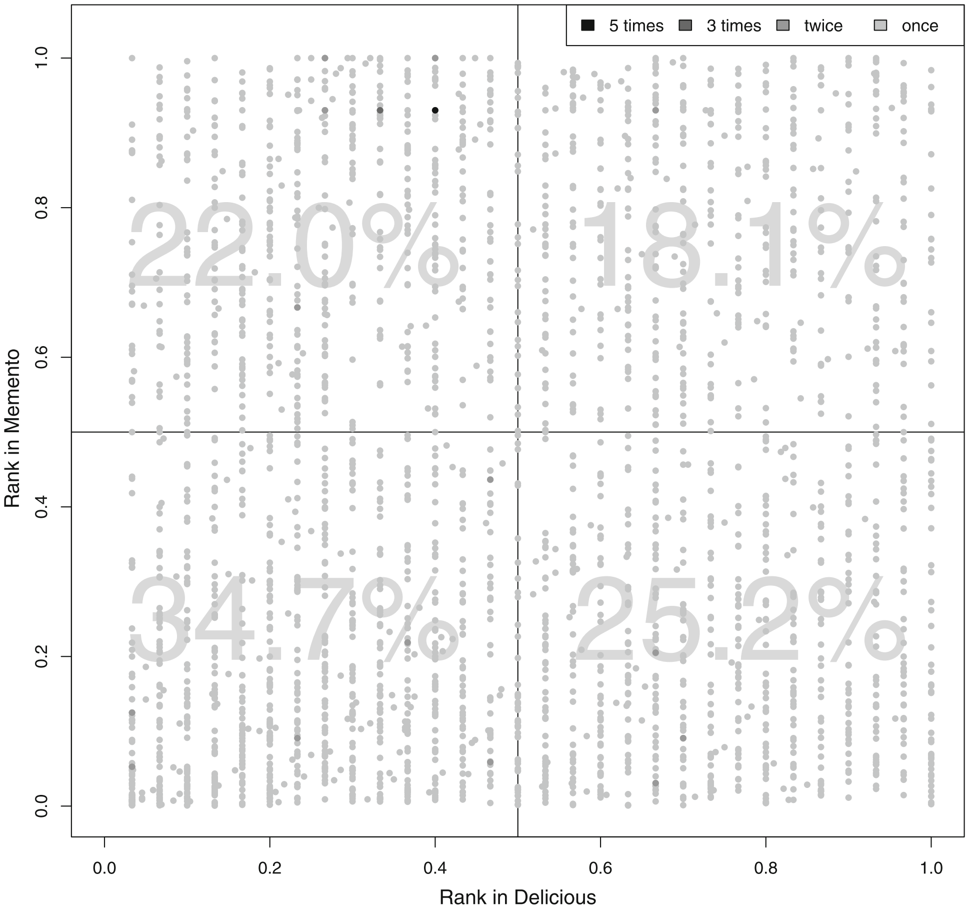


Fig. 8 Ghost tags ranks in delicious and corresponding Mementos

quadrant with a normalized Delicious rank and TF rank of ≤ 0.5. This indicates a high level of importance of the ghost tags for the document and also for the Delicious user. Further, one fourth of the ghost tags seemed to be more important for the document than in Delicious (second quadrant) and the inverse holds true for 22 % (third quadrant). In 18.1 % of all cases rather infrequently used terms became ghost tags.

The third result of this experiment is both the existence and the significance of ghost tags. One third of them were used very frequently in the document and very frequently used to annotate the page in Delicious.

5.4 Results

The series of experiments on the retrieval performance of tags has three main results:

1. Based on our corpus, we found that tags by themselves did not perform well and we did not find a significant difference for the tag-based query length in terms of number of tags. That is slightly disappointing since tags (if available) are rather easy to obtain and this method is applicable even if no Mementos of the missing Web page are available.
2. In combination with other methods, applying tag-based queries can improve the overall retrieval performance. As a secondary method tags performed similarly to lexical signatures but since they are easier to obtain, we promote the *T-TA-LS* sequence for rediscovering missing Web pages. However, the sequence is only applicable if tags are available for the missing URI. Since we have seen that tags are rather sparse, this sequence represents a best case scenario.
3. Ghost tags exist and they are significant. More than one-third of the ghost tags were used very frequently within the document and were very frequently used to annotate the page in Delicious.

6 Link neighborhood lexical signatures

It is well known that the link structure in the Web holds valuable information for search. Craswell et al. [72], for example, found that link anchor information can be more useful than the content itself for site finding. Dou et al. [73] provided indicators that anchor text is similar to user queries for search engines but also showed that anchors within the same site are less useful than external anchors. Kraft and Zien [74] propose the use of anchor text to refine search queries. They show that anchor text can provide terms for query refinement that perform better than terms obtained from a document's content itself.

In this section, we describe our experiments to investigate link neighborhood lexical signatures (LNLS) as a fourth method to rediscover missing Web pages. Just like tags, this method can also be applied even if no Mementos of missing pages are available. An LNLS of a Web page is a lexical signature generated from the content of other Web pages that link to the page of interest, also called their inlinks or backlinks. Since pages tend to link to related pages, our intuition was that the link neighborhood contains enough of the "aboutness" of the targeted page to create a well-performing search query. We tested several parameters to compute lexical signatures from those link neighborhoods to find the most effective signature-based implementation. We examined the effects of lexical signature size, backlink depth, and backlink ranking as well as the radius within a backlink page from which terms for the LNLSs were drawn.

6.1 Constructing the link neighborhood

We anticipated a large number of backlinks per URI, which made us use the same corpus of 309 URIs introduced in Sect. 4 for our experiment. For each URI, we queried the Yahoo! index to determine the pages that link to the URI ("backlinks"). The Yahoo! index has previously been shown to give more complete backlink results than other search engines [55]. 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 two. Figure 9 graphically explains such a link neighborhood. The page on the right (vertical lines) represents the target page with backlinks that is no longer available. In this example, we obtained three pages that link to the target page. These are the first-level backlinks, represented in the center with horizontal lines. We call the backlinks for the first-level backlinks second-level backlinks. They are represented with crossing lines. In this manner, we retrieved a total of 335, 334 pages, 28, 325 first-level and 306, 700 second-level backlink pages. For more detailed information about the generated link neighborhoods, we refer to our previous work [75,76].

6.2 Parameters of link neighborhood lexical signatures

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 were drawn. For every possible combination for each of these factors, we computed the TF–IDF value of every term in the appropriate section(s) of the appropriate pages. The LNLSs were generated based on the same simple TF–IDF method introduced earlier. Stop words were dismissed in advance but no stemming algorithms were applied. We utilized the Yahoo! BOSS API to obtain $|d_i|$ values for the LNLS computation.

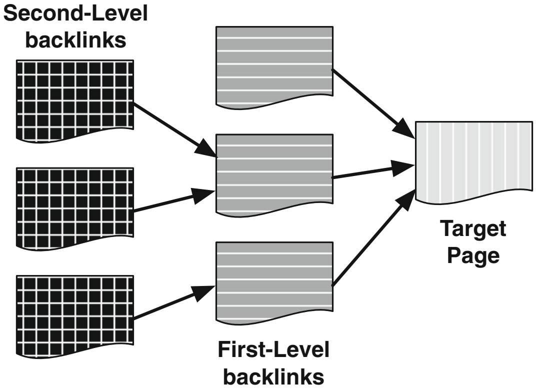


Fig. 9 Graphical example for a link neighborhood

Backlink Depth The two options for depth were:

1. to use the first-level backlinks only or
2. to use first- and second-level backlinks.

Our reasoning was that first-level backlinks might result in an LNLS that more accurately describes the missing page since they are closer to the target page. However, in cases where few first-level backlinks exist, second-level backlinks might provide more information, leading to a better performing LNLS.

Radius Lexical signatures are typically drawn from the entire page. 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 LNLS. 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 a "paragraph" of context around the link. We, therefore, considered the following four possibilities for the radius within the backlink page from which LNLSs were drawn:

1. from the entire page,
2. from the anchor text only,
3. from the anchor text ±5 terms, and
4. from the anchor text ±10 terms.

Backlink Ranking The backlinks returned from Yahoo! are ordered. To determine whether this ranking was helpful, we tested the following three possibilities:

1. using only the top 10 backlinks,
2. using the top 100 backlinks, and
3. using the top 1, 000 backlinks.

If fewer backlinks existed than allowed by the limit, we used all available backlinks. Our assumption was that if the rankings in backlink results were helpful, then using only the top backlinks would likely provide a better LNLS. If the ranking was not relevant, then using as many backlinks as possible might provide the better lexical signature since that would mean including more data.

LNLS Size We have previously shown that 5- and 7-term lexical signatures perform best. However, given that the lexical signatures in this experiment were derived from a link neighborhood instead of the target page itself, we needed to test the applicability of those parameters. We queried LNLSs of sizes one, two, three, four, five, six, seven, and ten.

6.3 Performance of link neighborhood lexical signatures

For the evaluation of our results, we used our four retrieval scenarios introduced earlier but also applied the normalized Discounted Cumulative Gain (nDCG). We set the relevance score to 1 for an exact match of the target URI, and 0 otherwise. We checked the first 100 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. Regardless of what parameters we set, we saw a dramatic decline in scores in all our experiments when we included second-level backlinks. This shows that second-level backlinks' relation to the target page was not tight enough to be useful in describing the target page. As our first result, we state that our best-performing method included only first-level backlinks.

With respect to the radius, we found that the anchor text only performed best. The performance with ±5 words or ±10 words added was equally bad and using the whole page performed the worst. Each step taken away from the anchor text, by broadening the radius to include words around the anchor or the entire page, yielded increasingly poor results. As our second result, we state that using the anchor text only performed best.

The analysis of the backlink ranks returned somewhat surprising results. The scores were very similar for either of the three options. However, using 1, 000 backlinks (and anchor text) showed the highest overall scores even though by a small margin. This constitutes our third result.

Table 8 Result rank and nDCG vs. lexical signature size (1-anchor-1,000)

Result rank	# of terms in lexical signature							
	1	2	3	4	5	6	7	10
1	32.11	50.50	**58.19**	54.85	52.51	45.82	38.80	23.41
2–10	10.03	10.70	7.02	5.35	2.34	2.34	1.67	0.33
11–100	5.69	3.34	0.67	0.33	0.33	0.33	0.33	0.67
> 100	53.41	36.79	35.45	40.80	46.15	52.84	60.53	76.92
Mean nDCG	0.38	0.56	**0.61**	0.57	0.53	0.47	0.39	0.23

The best values are in bold

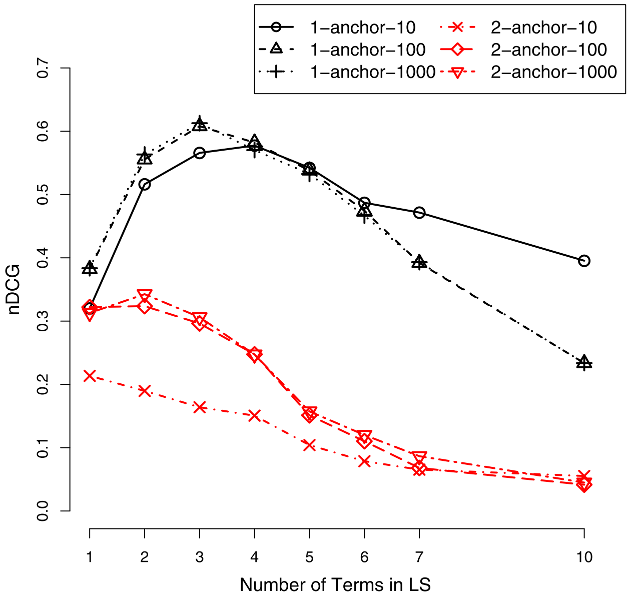


Fig. 10 First- and second-level Backlinks **Anchor** radius lexical signatures with various backlink ranks (shown as levels-radius-ranks)

The results of the experiments for the best performing LNLS in terms of its length were also intriguing since they diverged from what we have previously seen in Sect. 3. Table 8 shows the percentage of URIs in our four retrieval cases distinguished by length of the LNLS. The data were obtained using the best performing parameters, meaning the first level backlinks only, anchor text only and the top 1, 000 results regarding the backlink ranking. We can see that 3-term LNLSs performed best. They returned the most URIs top ranked and left the fewest undiscovered. This is unlike the results seen in Sect. 3. We consider the source of the terms that make up the LNLS to be the reason for this disparity. Here, the terms were 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 themselves, rather than the content of the target page. Using fewer terms, we decreased the risk of including a term in the lexical signature that did not appear in the target page.

Figure 10 provides an overview of our results. It shows average scores of methods based on anchor text, the first- and second-level backlinks and a variety of 10, 100, and 1, 000 backlink results included. First-level backlink methods were drawn in black and second-level methods in red. The x-axis is the number of terms included in the lexical signature and the

Table 9 Result rank and nDCG vs. lexical signature Size (1-anchor-10)

Result rank	# of terms in lexical signature							
	1	2	3	4	5	6	7	10
1	25.08	45.15	52.51	**55.85**	52.84	47.83	46.49	39.13
2–10	9.03	9.70	7.02	3.34	2.01	1.34	1.00	0.67
11–100	8.03	4.68	2.01	0.67	0.67	0.33	0.33	0.33
> 100	57.86	40.47	38.46	40.13	44.48	50.50	52.17	59.87
Mean nDCG	0.32	0.52	0.57	**0.58**	0.54	0.49	0.47	0.40

The best values are in bold

y-axis is the mean nDCG. The figure confirms the findings summarized in Table 8. 3-term LNLS with 1, 000 backlinks performed best. However, we can see that 4-term LNLSs with 10 backlinks performed fairly well also. Considering the huge implied cost to acquire ten or one hundred times as many pages and generate an LNLS based on an accordingly larger bucket of words, we were motivated to look further into results obtained with only the top 10 backlinks.

Table 9 shows our results for using the top 10 backlinks only. We can see that 4-term LNLSs performed well with almost 56 % URIs returned at the top position. Even though these numbers were not quite as good as the 3-term 1, 000 backlinks based LNLSs, given the implied costs, we consider the 4-term 10 backlink LNLSs preferable. This is the fourth finding of this experiment.

We also tested all logical combinations of previously introduced methods with LNLSs but did not find an improved retrieval performance. With this fact, plus the implied costs to generate LNLSs, we consider this method as a last resort, which could be applied if all other methods have failed.

6.4 Results

This method is applicable even if no Mementos of the missing page are available. However, LNLSs perform poorly and they are expensive to generate which means they should be considered a last resort for cases where all other options for a user have failed.

The series of experiments on the retrieval performance of LNLSs has four main results which are represented in our recommended parameters for the generation of LNLSs:

1. The inclusion of second-level backlinks only hurt the performance and hence only first-level backlinks are to be included.
2. Widening the radius to draw terms did not improve the performance. We recommend using the anchor text only for the bag of words to generate the LNLS.
3. We found that LNLSs based on 10 backlinks and
4. consisting of 4 terms are performing best, considering the costs involved to generate LNLSs.

7 Future work and conclusions

7.1 Future work

The here presented experiments were conducted over the period of five years and they are based on several Web page corpora of varying sizes. Intuitively, we see value in repeating these experiments within a shorter time span utilizing a single large and up-to-date corpus that may address the consequences of our corpus selection outlined in Sect. 1. A good candidate for such a corpus could be the ClueWeb12 [77] dataset that contains more than 730 million English language Web pages crawled in the first half of 2012. Alternatively, the crawl data from the Common Crawl Foundation [78] could also be considered a suitable corpus. It contains more than two billion Web pages collected in 2013.

Besides the corpus selection, we see several other aspects of future work. Some Web servers do not return a 404 response to requests for missing content. They either return a 200 response (meaning OK) with content telling the user that the requested page could not be found or they simply redirect with a 300-level response to a custom page or even the index page of the site. These scenarios are known as “soft 404s” (see Sect. 2.1.1) and have not been properly addressed in this work. An automatic detection of soft 404s would be desirable and, after detection, our here introduced methods can be applied to discover the desired content.

Stop words are usually dismissed before generating lexical signatures. We see an opportunity to identify stop words in anchor text. Examples for “stop anchors” could be “here”, “click here”, and the link URI itself. Identifying these stop anchors could lower the complexity of generating LNLSs.

All four here investigated methods have shown to contribute to rediscovering missing Web pages. Lexical signatures and titles (when obtained from Mementos) can perform well on their own, while tags (if available) only seem to contribute in combination with other methods. The value of LNLSs seems to be a last resort, if all other options have failed. We consider a Web service that applies these methods and helps users overcome link rot for future work. Such a service would use the Memento protocol to obtain old copies of now missing pages and the Delicious API to provide tags of the missing pages. It would return a list of alternative pages, which are obtained from applying all or a subset of our methods. It could further maintain a memory of references between missing pages and user’s picks for alternates to expedite the process on repeated requests.

As part of this work, a prototype of a Web browser plugin, called *Synchronicity*, was implemented that used all here described methods to rediscover missing Web pages. However, due to changing search engine API policies on one hand and Web browser technologies on the other, this extension requires continuous maintenance which exceeded the boundaries of this proof-of-concept implementation.

7.2 Conclusions

In this article, we compare four methods to rediscover missing Web pages based on their copies in Web archives, their user generated tags, and their in- and out-links. We present the results of multiple experiments investigating the retrieval performance of the methods individually as well as in combination. The experiments are based on various corpora, mostly containing URIs sampled from DMOZ. The results are depending on the availability of copies of pages in Web archives (Mementos) and of tags from social annotation services such as Delicious. The analysis of the results enables us to determine the parameters of the best performing methods.

First, we investigated lexical signatures of Mementos of missing Web pages. We found that 7- and 5-term lexical signatures performed best, depending on the retrieval goal. 7-term lexical signatures returned the most top ranked URIs and 5-term lexical signatures showed the best mean rank. We further showed that the Yahoo! BOSS API returned the best results in comparison to Google and MSN Live. Lexical signatures older than four to five years did not perform well when trying to rediscover the current version of a missing page.

Secondly, we investigated the retrieval performance of titles of Mementos of Web pages. We found that titles were at least as well performing as lexical signatures. Given the fact that titles are much easier to obtain and assuming that the Memento has a title, we consider them the preferable method for rediscovering missing Web pages. We also showed that the sequence of querying titles first and lexical signatures second can improve the retrieval performance.

Another part of this experiment was to investigate how titles change over time compared to the content of Web pages. We found titles to be much more stable than content, which supports our preference for titles over lexical signatures as the primary retrieval method.

The purpose of the third experiment was to analyze tags provided by users to annotate Web pages. We obtained the tags from the bookmarking service Delicious and found them to be performing poorly by themselves. This result was independent of the length of tag-based queries. However, we provided evidence that applying the tag-based method in combination with titles and lexical signatures can improve the overall retrieval performance. The drawback of this method is that tags for URIs are rather sparse. The provided results, therefore, represent a best case scenario. We further discovered the existence of what we call "ghost tags" as tags that describe previous versions of Web pages better than current ones. We provided indicators that ghost tags are of significance for both the user, which was annotating the page, as well as for the document in which they occur.

Our fourth experiment was aimed at investigating the parameters for the best performing link neighborhood lexical signatures (LNLS). We found that LNLSs generated from the top ten first-level backlink pages, based on the anchor text only, and containing four terms performed best. Since this method is based on the content of pages linking to missing pages, it is the most expensive one to generate. Hence we did not combine it with any of the previous methods and rather consider it a last resort for the rediscovery of missing Web pages.

Based on the assumption that Web content is rarely completely lost but often just moved from one location to another, we have provided four methods to support the rediscovery of missing content. These methods rely on the Memento framework and third party indexes such as Delicious and search engines. They can help to alleviate the link rot problem in the Web and contribute to a better browsing experience by reducing confrontations with frustrating 404s.

Acknowledgments This work was in part supported by the Library of Congress, the National Digital Information Infrastructure and Preservation Program (NDIIPP), and the National Science Foundation, grant IIS 0643784. The authors would like to thank the Computer Science Department and the College of Sciences at Old Dominion University for their support.This article is the result of the author's Ph.D. work while at Old Dominion University. A lot of credit goes to Michael L. Nelson for his eternal support and guidance throughout this work. Further thanks go to Herbert Van de Sompel from the Los Alamos National Laboratory for his valuable feedback to the article. Jeb Ware and Moustafa Emara deserve credit for contributing to this work during their studies at ODU. Lastly, we would like to thank the anonymous reviewers and editors of this article for their valuable feedback and stimulating discussion about sampling URIs from the Web.

References

1. Berners-Lee, T.: Cool URIs don't change http://www.w3.org/Provider/Style/URI.html (1998)
2. McCown, F., Marshall, C.C., Nelson, M.L.: Why websites are lost (and how they're sometimes found). Commun. ACM 52(11) (2008)
3. Van de Sompel, H., Nelson, M.L., Sanderson, R., Balakireva, L., Ainsworth, S., Shankar, H.: Memento: time travel for the web. Tech. Rep. arXiv:0911.1112 (2009)
4. Klein, M.: Using the Web Infrastructure for Real Time Recovery of Missing Web Pages. Ph.D. thesis, Old Dominion University (2011)
5. Henzinger, M.R., Heydon, A., Mitzenmacher, M., Najork, M.: On near-uniform URL sampling. Comput. Netw. **33**(1–6), 295–308 (2000)
6. Rusmevichientong, P., Pennock, D.M., Lawrence, S., Giles, C.L.: Methods for sampling pages uniformly from the world wide web.

In: AAAI Fall Symposium on Using Uncertainty Within Computation, pp. 121–128 (2001)

7. Harth, A., Umbrich, J., Decker, S.: MultiCrawler: a pipelined architecture for crawling and indexing semantic web data. In: The Semantic Web-ISWC 2006, vol. 4273, pp. 258–271 (2006)
8. Noll, M.G., Meinel, C.: Exploring social annotations for web document classification. In: Proceedings of SAC '08, pp. 2315–2320 (2008)
9. Umbrich, J., Harth, A., Hogan, A., Decker, S.: Four heuristics to guide structured content crawling. In: Proceedings of ICWE '08, pp. 196–202 (2008)
10. Klein, M.: The "Book of the Dead" Corpus. http://ws-dl.blogspot.com/2011/06/201-06-17-book-of-dead-corpus.html
11. Ainsworth, S.G., Alsum, A., SalahEldeen, H., Weigle, M.C., Nelson, M.L.: How much of the web is archived? In: Proceedings of JCDL '11, pp. 133–136 (2011)
12. Adar, E., Teevan, J., Dumais, S.T., Elsas, J.L.: The web changes everything: understanding the dynamics of web content. In: Proceedings of WSDM '09, pp. 282–291 (2009)
13. Cho, J., Garcia-Molina, H.: The evolution of the web and implications for an incremental crawler. In: Proceedings of VLDB '00, pp. 200–209 (2000)
14. Cho, J., Garcia-Molina, H.: Estimating frequency of change. ACM Trans. Internet Technol. **3**, 256–290 (2003)
15. Dalal, Z., Dash, S., Dave, P., Francisco-Revilla, L., Furuta, R., Karadkar, U., Shipman, F.: Managing distributed collections: evaluating web page changes, movement, and replacement. In: Proceedings of JCDL '04, pp. 160–168 (2004)
16. Fetterly, D., Manasse, M., Najork, M., Wiener, J.: A large-scale study of the evolution of web pages. In: Proceedings of WWW '03, pp. 669–678 (2003)
17. Lim, L., Wang, M., Padmanabhan, S., Vitter, J.S., Agarwal, R.C.: Characterizing web document change. In: Proceedings of WAIM '01, pp. 133–144 (2001)
18. Ntoulas, A., Cho, J., Olston, C.: What's new on the web?: the evolution of the web from a search engine perspective. In: Proceedings of WWW '04, pp. 1–12 (2004)
19. Ashman, H.: Electronic document addressing: dealing with change. ACM Comput. Surv. **32**(3), 201–212 (2000)
20. Ashman, H., Davis, H., Whitehead, J., Caughey, S.: Missing the 404: link integrity on the world wide web. In: Proceedings of WWW '98, pp. 761–762 (1998)
21. Davis, H.C.: Referential integrity of links in open hypermedia systems. In: Proceedings of HYPERTEXT '98, pp. 207–216 (1998)
22. Davis, H.C.: Hypertext Link Integrity. ACM Comput. Surv. **31** (1999). doi:10.1145/345966.346026
23. Johnson, D., Tanimoto, S.: Reusing web documents in tutorials with the current-documents assumption: automatic validation of updates. In: Proceedings of EDMEDIA'99, pp. 74–79 (1999)
24. Kahle, B.: Preserving the internet. Sci. Am. **276**, 82–83 (1997)
25. Lawrence, S., Pennock, D.M., Flake, G.W., Krovetz, R., Coetzee, F.M., Glover, E., Nielsen, F.A., Kruger, A., Giles, C.L.: Persistence of web references in scientific research. Computer **34**(2), 26–31 (2001)
26. Koehler, W.C.: Web page change and persistence—a four-year longitudinal study. J. Am. Soc. Inf. Sci. Technol. **53**(2), 162–171 (2002)
27. Spinellis, D.: The decay and failures of web references. Commun. ACM **46**(1), 71–77 (2003). doi:10.1145/602421.602422
28. Dellavalle, R.P., Hester, E.J., Heilig, L.F., Drake, A.L., Kuntzman, J.W., Graber, M., Schilling, L.M.: Information science: going, going, gone: lost internet references. Science **302**(5646), 787–788 (2003). doi:10.1126/science.1088234
29. McCown, F., Chan, S., Nelson, M.L., Bollen, J.: The availability and persistence of web references in D-Lib magazine. In: Proceedings of IWAW'05 (2005)
30. Nelson, M.L., Allen, B.D.: Object persistence and availability in digital libraries. D Lib Mag. **8**(1) (2002). doi:10.1045/january2002-nelson
31. Sanderson, R., Phillips, M., Van de Sompel, H.: Analyzing the persistence of referenced web resources with memento. In: Proceedings of OR '11 (2011)
32. Bar-Yossef, Z., Broder, A.Z., Kumar, R., Tomkins, A.: Sic transit gloria telae: towards an understanding of the web's decay. In: Proceedings of WWW '04, pp. 328–337 (2004)
33. Lee, T., Kim, J., Kim, J.W., Kim, S.R., Park, K.: Detecting soft errors by redirection classification. In: Proceedings of WWW '09, pp. 1119–1120 (2009)
34. Meneses, L., Furuta, R., Shipman, F.: Identifying "Soft 404" error pages: analyzing the lexical signatures of documents in distributed collections. In: Proceedings of TPDL' 12 (2012)
35. Fielding, R., Gettys, J., Mogul, J., Frystyk, H., Masinter, L., Leach, P., Berners-Lee, T.: Hypertext Transfer Protocol-HTTP/1.1 RFC-2612. Updated by RFC 2817 (1999)
36. Martinez-Romo, J., Araujo, L.: Recommendation system for automatic recovery of broken web links. In: Proceedings of IBERAMIA '08, pp. 302–311 (2008)
37. Martinez-Romo, J., Araujo, L.: Retrieving broken web links using an approach based on contextual information. In: Proceedings of HT '09, pp. 351–352 (2009)
38. Martinez-Romo, J., Araujo, L.: Analyzing information retrieval methods to recover broken web links. In: Proceedings of ECIR '10, pp. 26–37 (2010)
39. Francisco-Revilla, L., Shipman, F., Furuta, R., Karadkar, U., Arora, A.: Managing change on the web. In: Proceedings of JCDL '01, pp. 67–76 (2001)
40. Bogen, P., Pogue, D., Poursardar, F., Shipman, F., Furuta, R.: WPv4: A re-imagined Waldens paths to support diverse user communities. In: Proceedings of JCDL '11 (2011)
41. Harrison, T.L., Nelson, M.L.: Just-in-time recovery of missing web pages. In: Proceedings of HYPERTEXT '06, pp. 145–156 (2006)
42. Haslhofer, B., Popitsch, N.: DSNotify—detecting and fixing broken links in linked data sets. In: Proceedings of DEXA '09, pp. 89–93 (2009)
43. Popitsch, N.P., Haslhofer, B.: DSNotify: Handling broken links in the web of data. In: Proceedings of WWW '10, pp. 761–770 (2010)
44. Jones, K.S.: Index Term Weighting. Inf. Storage Retr. **9**(11), 619–633 (1973)
45. Park, S.T., Pennock, D.M., Giles, C.L., Krovetz, R.: Analysis of lexical signatures for improving information persistence on the world wide web. ACM Trans. Inf. Syst. **22**(4), 540–572 (2004). doi:10.1145/1028099.1028101
46. Baeza-Yates, R.A., Ribeiro-Neto, B.: Modern Information Retrieval. Addison-Wesley Longman Publishing Co., Inc, Boston (1999)
47. Frakes, W.B., Baeza-Yates, R.A. (eds.): Information Retrieval: Data Structures and Algorithms. Prentice-Hall, Englewood Clifs (1992)
48. Manning, C.D., Raghavan, P., Schtze, H.: Introduction to Information Retrieval. Cambridge University Press, New York (2008)
49. Robertson, S.E., Walker, S.: Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval. In: Proceedings of SIGIR '94, pp. 232–241 (1994)
50. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Inf. Process. Manag. **24**(5), 513–523 (1988). doi:10.1016/0306-4573(88)90021-0
51. Klein, M., Nelson, M.L.: A Comparison of techniques for estimating IDF values to generate lexical signatures for the web. In: Proceeding of WIDM '08, pp. 39–46 (2008)
52. The size of the World Wide Web. http://www.worldwidewebsize.com/

53. Phelps, T.A., Wilensky, R.: Robust Hyperlinks Cost Just Five Words Each. Tech. Rep. UCB//CSD-00-1091, University of California at Berkeley, Berkeley, CA, USA (2000)
54. Phelps, T.A., Wilensky, R.: Robust hyperlinks: cheap, everywhere, now. In: Proceedings of DDEP'00 (2000)
55. McCown, F., Nelson, M.L.: agreeing to disagree: search engines and their public interfaces. In: Proceedings of JCDL '07, pp. 309–318 (2007)
56. Agichtein, E., Zheng, Z.: Identifying "Best Bet" web search results by mining past user behavior. In: Proceedings of KDD '06, pp. 902–908 (2006)
57. Jansen, B.J., Spink, A., Saracevic, T.: Real life, real users, and real needs: a study and analysis of user queries on the web. Inf. Process. Manag. **36**(2), 207–227 (2000). doi:10.1016/S0306-4573(99)00056-4
58. Joachims, T., Granka, L., Pan, B., Hembrooke, H., Gay, G.: Accurately interpreting clickthrough data as implicit feedback. In: Proceedings of SIGIR '05, pp. 154–161 (2005)
59. Joachims, T., Granka, L., Pan, B., Hembrooke, H., Radlinski, F., Gay, G.: Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search. ACM Trans. Inf. Syst. **25**(2), 7 (2007). doi:10.1145/1229179.1229181
60. Klöckner, K., Wirschum, N., Jameson, A.: Depth- and breadth-first processing of search result lists. In: Proceedings of CHI '04, pp. 1539–1539 (2004)
61. Järvelin, K., Kekäläinen, J.: Cumulated gain-based evaluation of IR techniques. ACM Trans. Inf. Syst. **20**(4), 422–446 (2002)
62. Chakrabarti, D., Kumar, R., Punera, K.: Generating succinct titles for web URLs. In: Proceeding of KDD '08, pp. 79–87 (2008)
63. Klein, M., Shipman, J., Nelson, M.L.: Is this a good title? In: Proceedings of Hypertext '10, pp. 3–12 (2010)
64. Levenshtein, V.I.: Binary codes capable of correcting deletions. Inser. Reversals Soviet Physics Doklady **10**(8), 707–710 (1966)
65. Bao, S., Xue, G., Wu, X., Yu, Y., Fei, B., Su, Z.: Optimizing web search using social annotations. In: Proceedings of WWW '07, pp. 501–510 (2007)
66. Jason Morrison, P.: Tagging and searching: search retrieval effectiveness of folksonomies on the world wide web. Inf. Process. Manag. **44**(4), 1562–1579 (2008)
67. Heymann, P., Koutrika, G., Garcia-Molina, H.: Can social bookmarking improve web search? In: Proceedings of WSDM '08, pp. 195–206 (2008)
68. Bischoff, K., Firan, C., Nejdl, W., Paiu, R.: Can all tags be used for search? In: Proceedings of CIKM '08, pp. 193–202 (2008)
69. Delicious Integrated Into Yahoo Search Results. http://techcrunch.com/2008/01/19/delicious-integrated-into-yahoo-search-results/
70. Klein, M.: Adventures with the delicious API. http://ws-dl.blogspot.com/2011/03/2011-03-09-adventures-with-delicious.html
71. Klein, M., Nelson, M.L.: Evaluating methods to rediscover missing web pages from the web infrastructure. In: Proceedings of JCDL '10, pp. 59–68 (2010)
72. Craswell, N., Hawking, D., Robertson, S.: Effective site finding using link anchor information. In: Proceedings of SIGIR '01, pp. 250–257 (2001)
73. Dou, Z., Song, R., Nie, J.Y., Wen, J.R.: Using anchor texts with their hyperlink structure for web search. In: Proceedings of SIGIR '09, pp. 227–234 (2009)
74. Kraft, R., Zien, J.: Mining anchor text for query refinement. In: Proceedings of WWW '04, pp. 666–674 (2004)
75. Klein, M., Ware, J., Nelson, M.L.: Rediscovering missing web pages using link neighborhood lexical signatures. In: Proceedings of JCDL '11, pp. 137–140 (2011)
76. Ware, J., Klein, M., Nelson, M.L.: Rediscovering missing web pages using link neighborhood lexical signatures. Tech. Rep. arXiv:1102.0930v1, CS Department, Old Dominion University, Norfolk, Virginia, USA (2011)
77. The ClueWeb12 Dataset. http://lemurproject.org/clueweb12/
78. Common Crawl Foundation. http://commoncrawl.org/