Find, New, Copy, Web, Page - Tagging for the (Re-)Discovery of Web Pages

Martin Klein and Michael L. Nelson

Old Dominion University, Department of Computer Science
Norfolk VA 23529

{mklein, mln}@cs.odu.edu

Abstract. The World Wide Web has a very dynamic character with resources constantly disappearing and (re-)surfacing. A ubiquitous result is the "404 Page not Found" error as the request for missing web pages. We investigate tags obtained from Delicious for the purpose of rediscovering such missing web pages with the help of search engines. We determine the best performing tag based query length, quantify the relevance of the results and compare tags to retrieval methods based on a page's content. We find that tags are only useful in addition to content based methods. We further introduce the notion of "ghost tags", terms used as tags that do not occur in the current but did occur in a previous version of the web page. One third of these ghost tags are ranked high in Delicious and also occurred frequently in the document which indicates their importance to both the user and the content of the document.

1 Introduction

The World Wide Web is a highly dynamic information space where resources very frequently surface, disappear and move from one location to another. As one result users often encounter a "404 Page Not Found" response when requesting a web resource. This error may occur when re-visiting a bookmark created some time ago, requesting a no longer valid URI or following a link from a badly maintained web page. Even though we know how to create URIs that do not change [3] there are many reasons why URIs or even entire websites break [13].

In previous work [8, 9] we have provided two content based approaches to generate search engine queries that rediscover such missing pages. We are following our intuition that information on the web is rarely completely lost, it is just missing and we can utilize the web infrastructure (search engine, their caches, archives, etc) to rediscover and preserve these resources. Our previously introduced methods are the web page's title and the page's lexical signature. Both have shown to perform very well for the purpose of rediscovering missing web pages. However, both methods are applicable only if an old copy of the missing page can be found in the web infrastructure. If that fails we have no means to gain knowledge of the "aboutness" of the missing page.

As a third option we are motivated to investigate the retrieval performance of tags left by Delicious users to annotate URIs. We see several intriguing aspects for using tags: Unlike titles and lexical signatures tags may be available even if no old copy of a missing page can be found. That means even if we can not obtain the title or generate the lexical signature of the missing page we may find tags describing its content. Tags are created by many users, therefore somewhat utilize the "wisdom of the crowd". They have been predicted to be useful for search [4, 5] and shown to possibly contain terms that do not occur in the original (now missing) web page. This can be beneficial for retrieving other, potentially relevant documents. We do not expect tags to outperform titles and lexical signatures but we foresee an added value for the rediscovery of missing web pages in combination with the previously established methods. In previously generated corpora containing randomly sampled URIs we experienced that tags were very sparse. In [9] for example we only found tags for 15% of all URIs. This led us to the creation of a new, "tag-centric" corpus introduced here. In summary, this paper's contributions are:

- determining the best performing tag based query length in number of terms
- analyzing the similarity and relevance of tag based search results
- quantifying the increased retrieval performance for a combination of query methods
- identifying tags as ghosts of pages that have past.

2 Related Work

2.1 Tags for Search

A lot of work has been done to investigate the usefulness of tags for search. Morrison [6] for example found in an extensive study that search in folksonomies can be as precise as search in major modern web search engines. By comparing Delicious data with search engine log data Krause et al. [12] found that tags and search terms both follow a power law distribution. That implies a low overall overlap and an increased overlap for the more frequent terms. They further found sparse overlap in Delicious and search engine rankings but if there was overlap it occurred at the top end of the rankings. Heymann et al. [5] conducted the probably most extensive study on tags with a dataset of about 40 million bookmarks from Delicious. Their results show that about half of the tags occur in the content of the page they annotate and 16% even occur in the page's title. Interestingly they found that in one out of five cases the tags neither occur in the page nor in the page's in- or outlinks. They conjecture that tags therefore can provide data for search that is otherwise not available. However, they state that annotated URIs are rather sparse compared to the size of a modern search engine's index. Bischoff et al. [4] confirm the findings of Heymann et al. with almost 45% of their tags found in the page's text. Their user study shows that tags are mostly reliable and accurate and they are partially used the same way as search terms. Yanbe et al. [16] propose a social bookmarking-based ranking and use it to enhance existing link-based ranking methods. They also find that tag proportions stabilize over time which means users eventually come to an agreement over tags and even copy each other. The work done by Bao et al. [2] incorporates the frequency of tags used to annotate URIs as an indicator for its popularity and quality.

2.2 Content and Link Based Methods to Rediscover Web Pages

Content based search engine queries can be a powerful tool to rediscover missing web pages. We have shown in previous work [8] that lexical signatures are suitable. We found that 5− and 7−term lexical signatures perform best depending on whether the focus is on obtaining the highest mean rank (5 terms) or the most top ranked results (7 terms). Sugiyama et al. [14] have shown that the content of in- and outlinks of a web page can help refine the lexical signature of that page. We have built on that idea in [11] and determined optimal parameters to create a link neighborhood based lexical signature. Even though they are expensive to compute, similar to tags, they may provide an alternative if no copy of a missing page can be found in the web infrastructure. Further research in [9, 10] has shown that titles of web pages are a very strong alternative to lexical signatures. The results also prove that we can increase the retrieval performance by applying both methods combined.

3 Experiment Setup

3.1 Data Gathering

We have seen in previous work [9] that for datasets based on randomly sampled URIs tags are very sparse and it is hard to aggregate a somewhat representative corpus. Heymann et al. [5] supports this point by showing that compared to a search engine's index the number of URIs annotated with tags is diminishing. Therefore we decided to reverse the approach and obtain tags and the URIs they annotate instead of first sampling URIs and then asking for their tags hoping to get a good sized sample set. Note that these URIs are not really missing but due to the sparseness of tags we use the obtained URIs and pretend they are missing. A few sources are available to obtain tags left by users to annotate URIs. The website `delicious.com` is probably the most famous and most frequently used one. We queried Delicious for 5000 unique URIs from their index using the Delicious "random tool"[1]. We are aware of the bias of our dataset towards the Yahoo! index (which we query against) especially in the light of Yahoo! integrating Delicious data into their index[2]. However, sampling from Delicious is an approach taken by various researchers [4, 5].

We eventually aggregated 4968 unique URIs from Delicious. We did get 11 duplicates and despite the fact that we sampled from the "random tool" which

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

[2] `http://techcrunch.com/2008/01/19/delicious-integrated-into-yahoo-search-results/`

# of Tags	0	1-5	6-10	11-15	16-20	21-25	26-29	30
Frequency	0.42	1.44	2.36	4.48	6.66	6.86	4.04	**73.73**

Table 1. Tag Distribution

pulls from the Delicious index we obtained 21 URIs that did not have tags. We used screen scraping, instead of the Delicious API, to gather up to 30 tags per URI[3]. The order, which may be of relevance for web search, indicates the frequency of use for all tags. Table 1 shows the relative distribution by number of tags for all URIs. We obtain the maximum of 30 tags for almost three out of four URIs.

3.2 Performance Measure

We use the Yahoo! BOSS API for all queries and analyze the top 100 results. We apply three different performance measures for our evaluation. Since our data corpus consists of live URIs one way of judging the performance of tag based search queries is to analyze the result set and monitor the returned rank of the URI of interest. This establishes a binary relevance case. More precisely, similar to our evaluation in [9] the first performance measure distinguishes between four retrieval cases where the returned URI is:

1. top ranked
2. ranked 2-10
3. ranked 11-100
4. considered undiscovered (ranked 101+).

We consider URIs not returned within the top 100 as undiscovered. We are aware of the possibility of discriminating against results returned just above that threshold but it is known that the average user does not look past the first few search results ([1, 7]) which encourages our threshold. We also compute normalized Discounted Cumulative Gain (nDCG) for the result set as a measure to reward results at the top of the result set and penalize results at the lower end. We give a relevance score of 1 for an exact match of the target URI and a score of 0 otherwise. For comparison reasons we also include mean average precision (MAP) scores for our results with the same binary relevance scoring.

We secondly compute the Jaro-Winkler distance between the original URI and the top ten returned URIs from the result set. The intuition is that some highly relevant pages have very similar URIs. The Jaro-Winkler distance is frequently applied to measure the similarity between short string such as names. It is therefore well fitting for comparing our URIs.

As a third measure we compute the Dice coefficient between the content of the original page and the content of the top ten search results. This gives

[3] We have previously shown the Delicious API to be unreliable, see: `http://ws-dl.blogspot.com/2011/03/2011-03-09-adventures-with-delicious.html`

us a sense of the string based similarity between the original content and the returned results. A high coefficient means a high similarity which in turn can be interpreted as a high relevance to the query - the tags used to annotate the original URI.

4 Retrieval Performance of Tags

4.1 Length of Tag Based Search Queries

We determined the best performing lexical signature length in previous work [8] to be 5 and 7 terms and initially assumed these parameters could be equally applied to tags. Hence we created queries consisting of 5 and 7 tags and issued them against the API. It turns out our assumption was inaccurate and therefore we widened the spectrum. Table 2 shows query lengths varying from 4 to 10 tags and their performance in relative numbers with respect to our four retrieval categories introduced in Section 3.2 as well as their nDCG and MAP. The generally low mean nDCG and MAP values are due to the large number of undiscovered URIs. Table 2 shows that 8−tag queries return the most top ranked results (11%) and 7−tag queries, tied with 6−tag queries, leave the fewest URIs undiscovered. Is also shows that 7− and 8−tag queries are tied for the best mean nDCG while 8 tags have a slight edge at MAP. However, taking this data we can not find a statistical significance (p-value ≤ 0.05) between the performances of 5−, 6−, 7− and 8−tag queries. The performance of 4−, 9− and 10−tag queries is in comparison statistically significantly worse.

# of Tags	Top	Top10	Top100	Undis	Mean nDCG	MAP
4	7.2	11.3	9.6	71.9	0.14	0.11
5	9.0	11.3	**9.7**	69.7	0.16	0.13
6	9.7	**12.0**	9.0	**69.3**	0.17	0.14
7	10.5	11.5	8.7	**69.3**	**0.18**	0.14
8	**11.0**	10.8	8.1	70.1	**0.18**	**0.15**
9	10.3	9.9	8.0	71.9	0.17	0.14
10	9.7	8.9	6.4	75.0	0.15	0.13

Table 2. Relative Retrieval Numbers for Tag Based Query Lengths, nDCG and MAP

4.2 Relevance of Results

Our binary retrieval evaluation (the URI is either returned or not) is applicable since we know what the “right” result to the tag based query is - the URI. However, the results in Table 2 indicate that a large percentage of URIs remain undiscovered. We are now investigating the relevance and similarity of the returned results for cases where the URI of interest is not returned.

We compute the Jaro-Winkler distance between the original URI and the URIs of top ten results to determine the similarity between URIs. Given the data from Table 2 we take the results of the five best performing tag based query lengths (5, 6, 7 and 8 tags) for this analysis. Figure 1 shows in the left graph the mean Jaro-Winkler distance for all URIs (y-axis) per rank (x-axis). Even though the four lines differ by point character (with respect to their length) it seems insubstantial to distinguish between them. The mean Jaro-Winkler value is high. It varies between 0.59 and 0.62 with slightly higher values for the top two ranks. The values for ranks three through ten are almost indistinguishable. These results show very similar URIs in the top ten indicating a high degree of relevancy for the returned results.

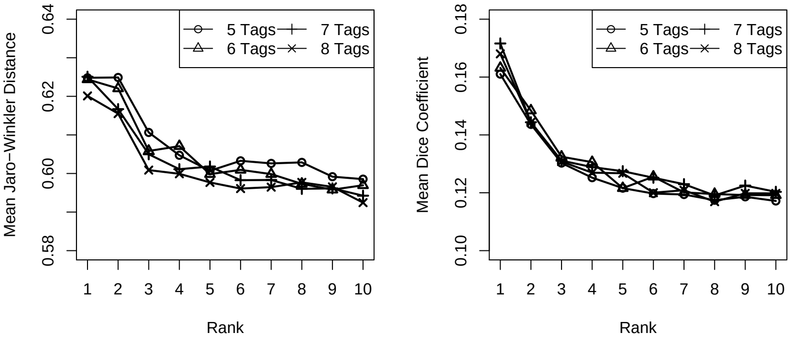


Fig. 1. Similarity Between URIs and Contents

Figure 1 shows in the right graph the Dice coefficient between the content of the URI the tags were derived from and the content of the top ten results. The intuition is that tags may not have the specificity to reliably return their URIs but contain enough information to return other relevant pages. This can especially be true for tags that do not actually occur in the pages. The graph also distinguishes by query length but the differences are diminishing. The mean Dice coefficient varies between 0.12 and 0.17. It is highest for the top two ranks and slightly decreases with higher ranks. The low mean Dice coefficients give an indication for a small degree of string similarity for the obtained results.

4.3 Performance Compared to Content Based Queries

In order to give a comparison for the performance of tags we also apply two content based methods. We extract the title of each page and generate its lexical signature. We issue our three queries (title, lexical signature, tags) for each URI against the API. Table 3 summarizes their performance distinguished by our four retrieval cases, nDCG and MAP. Note that the data in Table 3 is based

	Top	Top10	Top100	Undis	Mean nDCG	MAP
Titles	**60.2**	4.2	0.6	**34.9**	**0.63**	**0.62**
LSs	36.5	6.6	1.3	55.6	0.4	0.39
Tags	22.1	**15.4**	**10.2**	52.4	0.32	0.27

Table 3. Relative Retrieval Numbers for Titles, Lexical Signatures (LSs) and Tags, nDCG and MAP

on aggregated values meaning we merged the results for 5− and 7−term lexical signatures into one category and likewise for all tag based query lengths. We can see that titles outperform lexical signatures, supporting our earlier findings in [9, 10]. Both methods perform better than tags in terms of URIs returned top ranked, mean nDCG and MAP even though tags leave slightly fewer URIs undiscovered than lexical signatures. Tags return much more URIs in the top ten and top 100 than any other method. One interpretation of this observation is that tags, possibly rather generic by nature, are often not precise enough to return the URI top ranked. but they do provide enough specificity to return the pages within the top 100 results

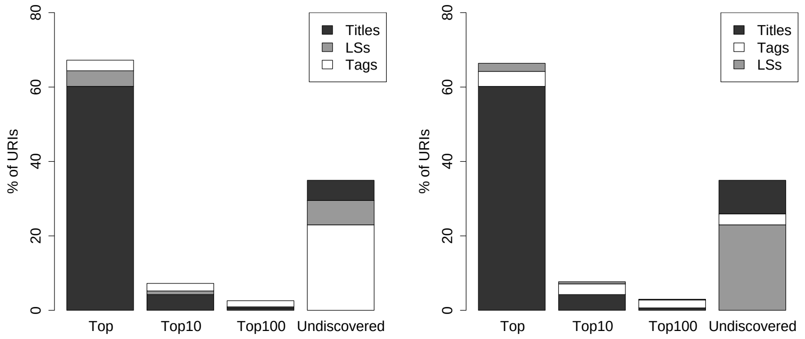
4.4 Combining Tags With Other Methods

Tables 2 and 3 show that the overall retrieval performance of tags alone is not very impressive. However what these tables do not show is the value of querying tags in combination with other methods. In other words, does the union of the results of more than one method improve the retrieval performance? And speaking from the preservation point of view, can we rediscover more missing URIs with combining two or even all three of the methods?

Extracting a web page's title from the content is cheap; it costs just one request to the resource. Lexical signatures are much more expensive to generate since each term, as a candidate to make it into the signature, requires the acquisition of a document frequency value. That means one request per unique term. Additionally we need to compute and potentially normalize term frequency (TF) values. Obtaining tags, similar to titles, is very cheap because it only requires one request to Delicious.

With this cost model in mind we define two combinations of methods: *Title-Lexical_Signature-Tags* (*T-LS-TA*) and *Title-Tags-Lexical_Signature* (*T-TA-LS*). Since titles perform best (as shown earlier and also demonstrated in previous work [9]) we maintain the priority for titles and query them as our first step in both combinations. As our second step in *T-LS-TA* we apply the lexical signature based method to all URIs that remained undiscovered (34.9% as shown in Table 3). We thirdly apply the tag based method to all URIs that are still undiscovered in *T-LS-TA*. The difference in the second combination is that we apply the tag based method second (to the 34.9%) and the lexical signature based method third.

Figure 2 shows the combined retrieval performance. The data of combination *T-LS-TA* is shown in Figure 2(a) distinguished by contribution per method and



(a) Title-Lexical Signature-Tag Combination

(b) Title-Tag-Lexical Signatures Combination

Fig. 2. Performance of Titles Combined with Lexical Signatures and Tags

separated in the previously introduced four retrieval categories. The first three bars (from left to right) are additive meaning 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 and the white part of the bars represents the URIs neither returned by titles nor by lexical signatures. They are returned 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 2(a) that 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 due to applying the second method. The height of the white portion of the bar corresponds to the final number of URIs that are left undiscovered after applying all three methods (23%) in the combination *T-LS-TA*. Figure 2(b) displays the data in the same way for the combination *T-TA-LS*. The color scheme remains 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 of course identical to the corresponding white bar in Figure 2(a). The additive bar for the top ranked results is slightly higher in Figure 2(a) (67.2% vs. 66.4%) but the bars for the top ten and top 100 results are slightly higher in Figure 2(b) (7.2% vs. 7.7% and 2.6% vs. 3.0%). The results for the combination of methods in terms of mean nDCG and MAP are summarized in Table 4. The performance increase of both combinations is statistically significant. Tags perform similarly compared to lexical signatures for URIs that remain undiscovered with the title method. Since tags are so much cheaper to obtain than lexical signatures these results lead to the recommendation to use tags a the default secondary method for

	TI	TI-LS	TI-LS-TA	TI-TA	TI-TA-LS
Mean nDCG	0.63	0.67	0.72	0.69	0.71
MAP	0.62	0.67	0.70	0.67	0.69

Table 4. Mean nDCG and Mean Average Precision for all Combinations of Methods

rediscovering missing web pages in case tags are available through Delicious. This condition is crucial since we have seen that tags were rather sparse for previously analyzed web page corpora.

5 Ghost Tags

Previous research [4, 5] has shown that about half the tags used to annotate URIs do not occur in the page's content. We find a slightly higher value with 66.3% of all tags not present in the page. If we consider the top ten tags only we find 51.5% of the tags not occurring in the page. This discrepancy intuitively makes sense since the ranking in Delicious is done by frequency of use which means that less frequently used tags are more likely to not appear in the page. However, these numbers only apply for the current version of the page. The tags provided by Delicious on the other hand are aggregated over an unknown period of time. The date of tags in Delicious can only be approximated but not reliably computed. It is possible that some tags used to occur in a previous version of the page and were removed or replaced at some point but still are available for that page through Delicious. We call these "ghost tags", terms that persist as tags after disappearing from the document itself.

To further investigate this aspect we use the Memento framework [15] to obtain old copies for all URIs that have tags not occurring in their content. For our dataset that applies to more than 95% of the URIs. Memento provides a timemap with references to all available Mementos (particular copies of a page at a certain point in time) per URI. Since we obtain 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 obtain Mementos of 3,306 URIs some of which date back to 1996. We find a total of 4.9% ghost tags. They occur in about one third of the previous versions of our web pages. Figure 3 displays the distribution of tags (dark gray) and the Mementos they occur in (light gray) per year. Almost the entire body of our "ghost tags" is found in Mementos from recent years. Both the most amount of ghost tags and their Mementos date back to 2009. We also see noticeable numbers from 2011 which indicates a very short time between the publication of the tags at which time they did occur in the page and their disappearance from the page. The majority of these very recent Mementos were obtained from search engine caches. The observations from Figure 3 confirm for one that ghost tags exist meaning some tags better represent the past content of a web page than the current and for two these ghost tags are found in the more recent past and rarely date back more than three years.

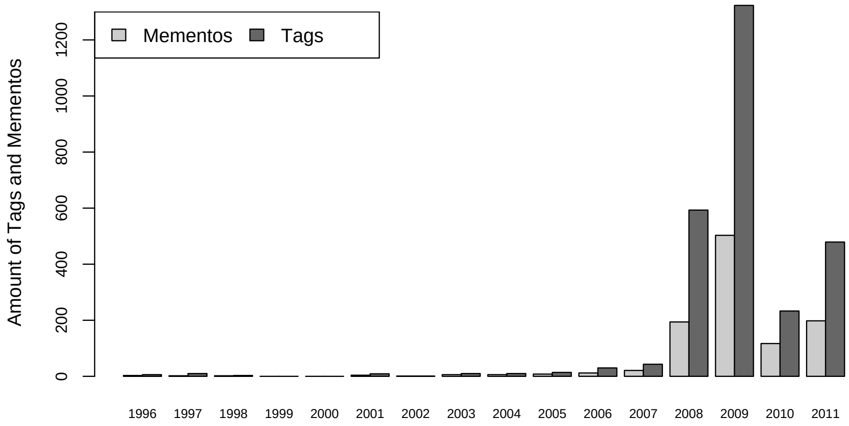


Fig. 3. Amount of Ghost Tags Occurring in Previous Versions of Web Pages

We then determine the importance of ghost tags for a page. We compare the tags' occurrence frequency in Delicious and their term frequency (TF) in the first available Mementos. We rank each ghost tag according to its Delicious and its TF rank and normalize the rank to a value between zero and one in order to avoid a bias towards a greater amount of available tags and longer documents. The closer the value gets to zero the higher the rank and the greater the importance. Figure 4 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 three times and one five times). The semi-transparent numbers indicate the percentage of dots or ghost tags in the corresponding quadrants. The numbers confirm our first visual impression of the graph. A majority of ghost tags (34.7%) occur in the first quadrant meaning their normalized Delicious rank is ≤ 0.5 and so is their TF rank. This indicates a high level of importance of the ghost tags for the document and also for the Delicious user. One fourth of the ghost tags seem to be more important for the document than in Delicious since their ranking there is > 0.5. On the other hand for 22% of ghost tags the inverse holds true. In 18.1% of the cases we can claim that "only" infrequent terms became ghost tags. These results show the significance of ghost tags since one third of them were used very frequently in the document and still are used frequently in Delicious.

6 Conclusions and Future Work

In this paper we have investigated the performance of tags for the purpose of discovering missing web pages. We obtained tags of almost 5,000 URIs from Delicious and showed that a search engine query containing five to eight tags performs best. More than 20% of the URIs are returned in the top ten ranks. We have further provided evidence for the top ten results to be similar to the URI the queried tags were obtained from. Compared to querying the title of the page or its lexical signature tags do not perform well but a combination of these methods increases the overall retrieval performance. We have also explored the notion of "ghost tags" as terms from Delicious that do not occur in the current version but do occur in a previous version of the web page. More than one out of three ghost tags appear to be important for the user as well as for the document since they rank high in Delicious and occur frequently in the text.

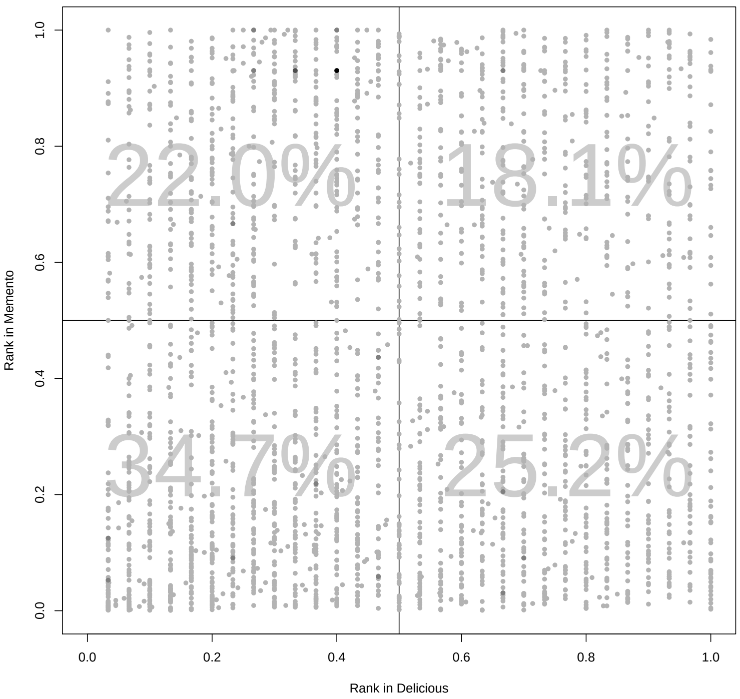


Fig. 4. Ghost Tags Ranks in Delicious and Corresponding Mementos

Our notion of ghost tags refers to the earliest available copy of web pages only. We will further investigate the aspect of time by including more copies of pages over time giving us a more precise idea of the age of the ghost tags. Another unanswered question is whether some tags predate the actual web page. If we can timestamp tags and monitor their frequency of use we can give a more specific description of their dynamics. In other words do users stop using tags when they disappear from or appear in the page? We have shown in previous work that titles and lexical signatures of web pages change over time. Naturally these methods after some time become obsolete as search engine queries. The question remains whether tags, as user given keywords, must be seen as dated at some point as well.

References

1. E. Agichtein and Z. Zheng. Identifying "Best Bet" Web Search Results by Mining Past User Behavior. In *Proceedings of KDD '06*, pages 902–908, 2006.
2. S. Bao, G. Xue, X. Wu, Y. Yu, B. Fei, and Z. Su. Optimizing Web Search Using Social Annotations. In *Proceedings of WWW '07*, pages 501–510, 2007.
3. T. Berners-Lee. Cool URIs don't change. 1998. `http://www.w3.org/Provider/Style/URI.html`.
4. K. Bischoff, C. Firan, W. Nejdl, and R. Paiu. Can All Tags Be Used for Search? In *Proceedings of CIKM '08*, pages 193–202, 2008.
5. P. Heymann, G. Koutrika, and H. Garcia-Molina. Can Social Bookmarking Improve Web Search? In *Proceedings of WSDM '08*, pages 195–206, 2008.
6. P. Jason Morrison. Tagging and Searching: Search Retrieval Effectiveness of Folksonomies on the World Wide Web. *Information Processing and Management*, 44:1562–1579, July 2008.
7. T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay. Accurately Interpreting Clickthrough Data as Implicit Feedback. In *Proceedings of SIGIR '05*, pages 154–161, 2005.
8. M. Klein and M. L. Nelson. Revisiting Lexical Signatures to (Re-)Discover Web Pages. In *Proceedings of ECDL '08*, pages 371–382, 2008.
9. M. Klein and M. L. Nelson. Evaluating Methods to Rediscover Missing Web Pages from the Web Infrastructure. In *Proceedings of JCDL '10*, pages 59–68, 2010.
10. M. Klein, J. Shipman, and M. L. Nelson. Is This a Good title? In *Proceedings of Hypertext '10*, pages 3–12, 2010.
11. M. Klein, J. Ware, and M. L. Nelson. Rediscovering Missing Web Pages Using Link Neighborhood Lexical Signatures. In *Proceedings of JCDL '11*, 2011.
12. B. Krause, A. Hotho, and G. Stumme. A Comparison of Social Bookmarking with Traditional Search. In *Proceedings of ECIR '08*, pages 101–113, 2008.
13. C. C. Marshall, F. McCown, and M. L. Nelson. Evaluating personal archiving strategies for internet-based information, 2007.
14. K. Sugiyama, K. Hatano, M. Yoshikawa, and S. Uemura. Refinement of TF-IDF Schemes for Web Pages using their Hyperlinked Neighboring Pages. In *Proceedings of HYPERTEXT '03*, pages 198–207, 2003.
15. H. Van de Sompel, M. L. Nelson, R. Sanderson, L. Balakireva, S. Ainsworth, and H. Shankar. Memento: Time Travel for the Web. Technical Report arXiv:0911.1112, 2009.
16. Y. Yanbe, A. Jatowt, S. Nakamura, and K. Tanaka. Can Social Bookmarking Enhance Search in the Web? In *Proceedings of JCDL '07*, pages 107–116, 2007.