The Metadata Triumvirate: Social Annotations, Anchor Texts and Search Queries

Michael G. Noll, Christoph Meinel
Hasso-Plattner-Institut an der Universität Potsdam, Germany
{michael.noll, meinel}@hpi.uni-potsdam.de

Abstract

In this paper, we study and compare three different but related types of "metadata" about web documents: social annotations provided by readers of web documents, hyperlink anchor text provided by authors of web documents, and search queries of users trying to find web documents. We introduce a large research data set called CABS120k08 which we have created for this study from a variety of information sources such as AOL500k, the Open Directory Project, del.icio.us/Yahoo!, Google and the WWW in general. We use this data set to investigate several characteristics of said metadata including length, novelty, diversity, and similarity and discuss theoretical and practical implications.

1 Introduction

The recent emergence and success of folksonomies and so-called *tagging* with services such as del.icio.us or Flickr have shown the great potential of this simple yet powerful approach to collect metadata about objects. Unlike traditional categorization systems, the process of tagging is nothing more than annotating documents with a flat, unstructured list of keywords called *tags*. The social aspect comes from sharing these annotations with other users. The number of peer-reviewed research on tagging has been increasing lately, and several studies have already analyzed the semantic aspects of tagging and why it is so popular and successful in practice [11], [18]. However, the scientific community is still in the process of discovering all facets of the information provided by social annotations [12, 20, 27].

On the other hand, indexing and ranking techniques in the information retrieval area have long been using incoming or outgoing hyperlinks of a web document to infer information about the document and its neighbors, for instance by associating terms with the web documents that are not part of the documents themselves [13, 7]. Here, the descriptive "annotations" such as incoming anchor texts help to gain more knowledge about the documents at hand, thereby leveraging the link structure of the WWW not only for ranking [14, 6] but also for indexing purposes. A third source of information are search query logs. A variety of concepts and techniques have been proposed to leverage information extracted from query logs, for example for classification of web queries [5], re-ranking of search results [29] or extracting semantics [3].

In this paper, we analyze and compare these three different but related types of "metadata" about web documents: social annotations provided by readers of web documents, anchor text of incoming hyperlinks provided by authors of web documents, and search queries of users trying to find web documents. We introduce a research data corpus called *CABS120k08* which contains a large set of web documents including information about their *C*ategorization, incoming *A*nchor text, social *B*ookmarks, and *S*earches. We use CABS120k08 to investigate how much information said metadata types provide, how they relate to each other, and what they can be used for.

The rest of the paper is organized as follows: In the next section, we describe the methodology and the data set for conducting the study. The experimental results are discussed in section 3. We summarize our findings in section 4 and describe related work in section 5.

2 Methodology

The analyses in this paper require access to a large and diverse collection of data. Web companies such as search engines or social sites conduct a number of internal analyses, however they are usually reluctant to publish specific results or research data sets. This can be for competitive reasons or to protect the privacy of their users. As a result, we had to build our own data set called *CABS120k08* by using publicly available information.

2.1 Data sampling

The first task is to create a large and representative set of web documents. Since we also want to study each metadata

type with regard to classification, all of these web documents need to be properly categorized, i.e. there should be a ground truth for categorization. We therefore decided to bootstrap our data set by combing data from the AOL500k collection and the Open Directory Project as follows.

The randomly sampled AOL500k corpus is one of the largest publicly available collections of search queries today [22]. It consists of 20 million web queries collected from 650,000 users on AOL Search over three months in 2006. As the result of these search queries, about 1.6 million different web documents where visited by users. What makes this collection so interesting for researchers is the vast amount of data and that the data "represents real world users, un-edited and randomly sampled" [22]. However, the publication of AOL500k was accompanied with privacy concerns. We have therefore discarded any user IDs during data sampling.

The Open Directory Project (http://www.dmoz.org/) is by its own account "the largest, most comprehensive human-edited directory of the Web". At the time we built our data set, the Open Directory contained 4,818,944 web documents in about 590,000 categories. The directory is constructed and maintained by a global community of volunteer editors which evaluate and categorize web documents into one or more predefined categories based on common standards and best practices to ensure consistency. In this work, we used Open Directory data for the categorization dimension in our data set and as the ground truth for classification analysis.

We built the initial set of web documents for CABS120k08 by an intersection of AOL500k and the Open Directory. Only such documents made into our data set which were both searched for and subsequently visited (AOL500k) as well as categorized (Open Directory). We also removed any such documents which were repeatedly unretrievable from the WWW. The resulting collection consists of 117,434 web documents, representing 7.3% and 2.5% of web documents originally in AOL500k and the Open Directory, respectively.

2.2 Data collection

The next task is to augment the data set with additional information for social annotations (by web readers), anchor texts from incoming hyperlinks (by web authors) and document popularity. We included document popularity because we found that user behavior for social bookmarking varies considerably with document popularity [19, 21].

For the social annotation dimension, we used del.icio.us, the main social bookmarking site in the WWW today. The Yahoo! company has a large user base with more than 2 million registered users in 2007. Using the popular del.icio.us also makes it easier to compare our results with other research publications based on del.icio.us data. For a complete picture of a web document in our study, we require the full history of its social bookmarks. Unfortunately, this type of data is not readily available from del.icio.us' official data channels such as its API or data feeds. We therefore developed a distributed parallel crawler on top of the open source Hadoop platform, which implements the concepts of the MapReduce framework [8]. For each document in our data set, the crawler extracted its full bookmarking history from del.icio.us' HTML web pages.

For anchor texts of incoming hyperlinks of a document, we built another distributed parallel crawler. A crawler instance would query Google for any known incoming hyperlinks of a given web document, also called its *inlinks* or *backlinks*. It would then download the referring pages from the WWW and extract the anchor texts of those hyperlinks pointing to the original web document. We used the anchor text definition of Kraft and Zien [15] which define anchor text as the text that appears within the bounds of an `HTML a` tag[1]. Due to technical reasons, we only processed up to 100 referring pages per document.

For measuring a web document's popularity in the WWW, we decided to use Google PageRank [6], partly because it is well studied in literature but also because of the cooperation of Google and AOL Search. The PageRank algorithm analyzes the WWW link structure, and is thus related to the analysis of a document's inlink anchor texts (provided by web authors) but not related to social annotations as seen on del.icio.us (provided by web readers). However, we have found in previous studies [19, 21] that the PageRank algorithm is still a good indication of whether a given web document is also socially popular. We therefore argue that a document's general popularity is measured reasonably well by its PageRank for the context of this work.

In the last step, we retrieved each document's source from the WWW. The final corpus is available for download from the first author's home page[2].

3 Experimental results

3.1 Overview

The total CABS120k08 data set consists of 117,434 web documents. By definition, each document in the set has been searched for at least once and is categorized into at least one category. We discarded the special tags `system:unfiled`' and `imported` from our analysis[3]

[1]For example, the anchor text of the hyperlink `<a href="labs.html">quux</a>` is "quux" and associated with the page `labs.html`.

[2]CABS120k08 is available for download at http://www.michael-noll.com/cabs120k08/

[3]Bookmarks without tags are auto-tagged by del.icio.us with `system:unfiled` whereas the tag `imported` is by default automat-

but left them in the data set as-is. An overview is shown in table 1. Estimated probabilities based on our observations are shown in table 2.

Total documents	117,434	
Total categories	84,663	
Total searches	2,617,326	
Total anchor texts	2,242,621	
Total users	388,963	
Total bookmarks	1,289,563	unique: 9.1 %
Total tags	3,383,571	unique: 26.3 %
Categorized documents*	117,434	100.0 %
Searched documents*	117,434	100.0 %
Anchored documents	95,230	81.1 %
Bookmarked documents	59,126	50.3 %
Tagged documents	56,457	48.1 %

**100% by definition*

Table 1. Overall statistics of CABS120k08.

A first surprise is the rather high probability of a document being bookmarked, 50.3%, compared to 81.1% for having at least one incoming hyperlink with anchor text. This is particularly interesting since del.icio.us - from which we extracted social bookmarking data - has been founded just five years ago in 2003 whereas the WWW - from whose link structure we derived anchor text data - has been around for much longer. Following the assumption that incoming hyperlinks are an indication that a web document is perceived as "interesting" or "important" by the referring party [6], bookmarks even cover $P(bookmarked \mid anchor\ text) = 57.5\%$ of relevant web documents in the corpus.

On the other hand, bookmarks by web readers are like incoming hyperlinks an indication that a web document is interesting [1]. In the data set, 3.7% of web documents have been bookmarked but do not have any incoming hyperlinks with anchor text. Since incoming hyperlinks are most likely created by human web authors, this observation indicates that these 3.7% may be interesting pages which have not been discovered by web authors yet. This result is similar to [12] which found that social bookmarking can serve as a small source for new, unindexed web pages, i.e. pages which have not been discovered by automated web spiders[4].

The incentives to add additional metadata to shared bookmarks, particularly via tags, have been found to go beyond selfish and organizational reasons such as refindability or quick access to web resources [1]. They are enhanced by social aspects such as recommendation or collaboration [2]. Our findings seem to confirm this: we can observe a strong tendency of users to add tags to their bookmarks: 95.5% of bookmarks have tags.

ically added to bookmarks during batch import by del.icio.us.

[4]Note that all web documents in the corpus have been extracted from AOL500k during data sampling (see section 2.1), meaning they have been returned as a search result by AOL Search and therefore must have been indexed at least by this search engine. We therefore cannot conclude based on evaluations of the CABS120k08 data set alone whether and by how much social bookmarks by human users can yield *unindexed* web pages, i.e. those pages that have not yet been discovered by automated web spiders.

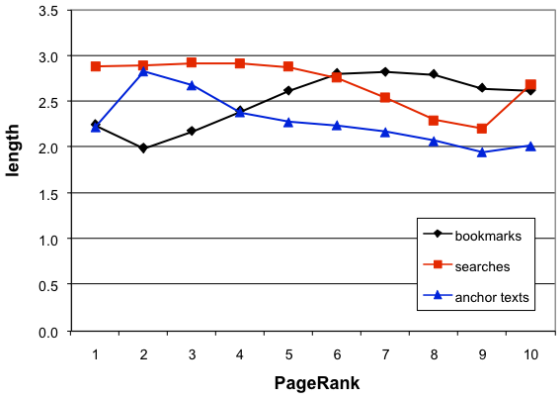

Figure 1. *Length* Average length of bookmarks, anchor texts and search queries.

$P(bookmarked \cap anchor\ text)$	0.467
$P(tagged \cap anchor\ text)$	0.447
$P(bookmarked \mid anchor\ text)$	0.575
$P(tagged \mid anchor\ text)$	0.552
$P(anchor\ text \mid bookmarked)$	0.927
$P(anchor\ text \mid tagged)$	0.930

Table 2. Estimated probabilities.

3.2 Length

The length of a search query, i.e. the number of keywords per query, has been studied in the past and reported as being rather short with 2.*x* keywords on average [25]. We were interested in comparing the length of search queries with the "length" of social bookmarks and anchor texts. We define the *length* of a bookmark to be the number of its tags, and the length of an anchor texts as the number of its words.

Globally, the average length of searches, bookmarks and anchor texts in the CABS120k08 data set are 2.89, 2.49 and 2.43, respectively. While it may seem at first glance that the length of bookmarks and anchor texts are almost equal, we found that the lengths vary significantly by document popularity as shown in figure 1. There are strong negative correlations with document popularity for search queries and anchor texts: Spearman-r are -0.82 and -0.81, respectively. On the other hard, there is a positive correlation with doc-

ument popularity for bookmarks: Spearman-r is +0.67[5]. This means in practice that anchor texts provide a larger amount of data for less popular web documents whereas social bookmarks do so for popular web pages, the break-even point being at PageRank 4. In our data set, the amount of data provided by the average anchor text is larger for 37% of documents (PR0-3) compared to 29% of documents in the case of social bookmarks (PR5-10). So in direct comparison, anchor texts "win" in the first $\frac{1}{3}$ of the cases, draw in the second $\frac{1}{3}$, and lose in last $\frac{1}{3}$; vice versa for bookmarks.

Looking at searches, the average search query dominates anchor text across all PageRanks. Compared to social bookmarks, search query length has a break-even point with bookmarks at PR6, and a second at PR10[6]. Here, the average search query provides more data for 90% of documents compared to 3% for social bookmarks.

Finally, we observe that the average lengths of all three metadata types stay between 2 and 3 terms even when taking variations due to document popularity into account. While we find it difficult to explain, users seem to prefer using only 2 or 3 terms per action even across different problem domains (social bookmarking, hyperlink creation, searching the web).

3.3 Novelty

A lot of tasks in information retrieval employ techniques to extract data from web documents, for example for indexing or classification purposes. On the other hand, not all information is captured by the terms in document content. Without further tricks such as anchor text analysis or latent semantic indexing [9], a web search for "biology" would not turn up any documents where the term "biology" would not appear in the document content [14].

In this section, we analyze how much "new" data is provided by social bookmarks, anchor texts and search queries. We are interested in finding out how much each metadata type is suited to add new information to web documents and thus how much it could help to improve information retrieval tasks in the context described above.

We define *novelty* as the percentage of unique terms which are not already present in a document. The terms for social bookmarks are represented by the set of unique tags aggregated over all bookmarks of a document, i.e. if multiple users add the tag `research`, it is counted only once. The terms for anchor texts (unique words) and search queries (unique keywords) are defined similarly. The corresponding document is represented by the set of unique words in its content, which we define as the sum of its `HEAD` `title`, `META keywords`, `META description` and `BODY`. Details are shown in figure 2.

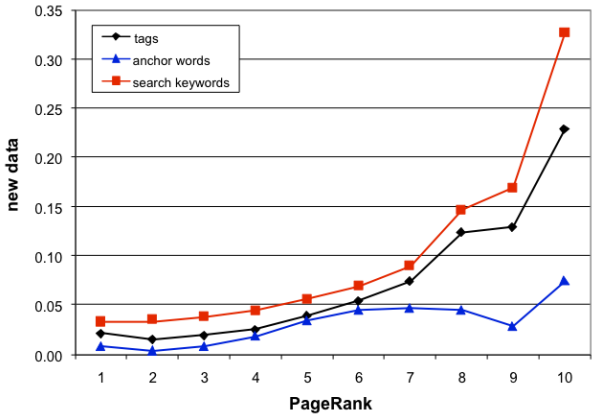

Figure 2. *Novelty* Percentage of new data provided by a document's tags, anchor words and search keywords. For example, 7.5 % of tags of a PR7 document are not present in the document's content.

Firstly, we observe that the amount of new information is comparatively low. All three types of data stay below 6% novelty for about 90% of documents in our data set (PR0-5). Search keywords dominate tags which in turn dominate anchor text words. Interestingly, the curves of search keywords and tags show similar behavior: both increase with document popularity with larger increases starting at PR6. Novelty for words in anchor text basically stays below the 5% threshold and peaks at PR7 and PR10[7].

Secondly, we find that tags provide more new data than anchor texts. This indicates that tags are a better source for new data, particularly for popular web documents. However, we will see in section 3.5 that the similarity of tags and anchor texts is relatively low, which indicates that they provide different kinds of information. We therefore argue that if one is interested in identifying new data, one should consider trying both metadata types.

We conclude that the majority of available metadata provided by social bookmarks, anchor texts or search queries does add only a small amount of new information to web documents. This observation is similar to the results of [12] which found that a search engine is unlikely to gain much from tags of social bookmarks if it already has access to document content.

[5] Kendall-τ for search queries, anchor texts and social bookmarks are -0.64, -0.73 and +0.47, respectively.

[6] The interpretation of the break-even point at PR10 should be treated with care since our data set contains only five web documents with PR10.

[7] As we said previously, PR10 results should be treated with care since there are only five documents with PR10 in the data set.

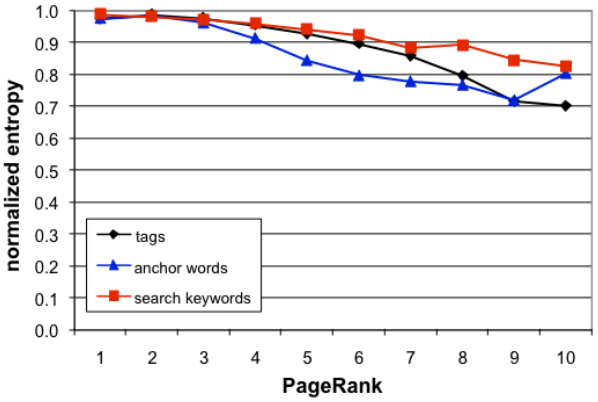

Figure 3. *Diversity* Normalized entropy of tags, anchor words and search keywords. A value of 0 denotes zero entropy (uniformity), a value of 1 maximum entropy (high diversity).

3.4 Diversity

In this section, we study the inherent diversity of information provided by social bookmarks, anchor texts and search queries. Generally, we can assume that users do not collaborate when searching the Internet or creating web documents with hyperlinks and anchor texts to other pages. There is a collaboration aspect for social bookmarking and tagging, but it is one facet of many [18, 2]. In a previous work [21], we have used entropy to measure the diversity of social annotations. A document's tags and their tag counts can be considered as a "tag histogram", and the entropy E of such an histogram can be computed by

$$E(d) = -\sum_{t_i \in T(d)} p(t_i|d) \log_2 p(t_i|d) \qquad (1)$$

where $T(d)$ is the set of tags with which document d has been annotated and $p(t_i|d)$ is the probability of d being annotated with tag t_i. We use the observed frequencies in our data set to estimate the probabilities $p(t_i|d)$. We define similar entropies for anchor texts (words and their counts) and search queries (keywords and their counts). We normalize the entropies so that zero entropy is represented by 0 and maximum entropy by 1. The results are shown in figure 3.

Firstly, there are strong negative correlations with document popularity for all metadata types: Spearman-r for tags, anchor words and search keywords are -0.96, -0.87 and -0.99, respectively[8]. With increasing document popularity, the diversity of information decreases.

Secondly, search queries show the highest diversity. The reason could be that searching the Internet is arguably the most "random" user action in our study. In contrast, users create bookmarks or hyperlinks with anchor text only *after* reading a document *and* perceiving it as useful. This process seems to serve as a kind of "noise filter" which search queries are lacking. Similarly, users do not only have problems with finding relevant information in the WWW per se, they also have problems with formulating good search queries [24]. Additional effects such as users becoming accustomed to automatic spell correction by search engines might further increase the diversity for search queries.

Thirdly, tags are generally more diverse than anchor texts. On one hand, this result suggests that tags are noisier than anchor texts and therefore potentially less useful. On the other hand, Bao et al. [4] observed that tags provide a multi-faceted summary of web documents. Seen this way, the diversity of tags could be an advantage since it might capture information and meanings that anchor texts miss. Additionally, we found [21] that tag noise does indeed provide relevant data for information retrieval and classification tasks. These results suggest that tags do provide valuable information but it is important to separate signal from noise. A simple way to do so is applying thresholding or considering only the top n tags, a technique commonly used for creating the so-called *tag clouds*. A more sophisticated way is to study the structure and dynamics of social networks for identifying expert users and expert user groups, thus adding a trust layer on top of social annotations. Interestingly, this is related to the analysis of WWW structure for identifying link farm spam pages [26]. We are preparing such kinds of trust analyses as part of our future research.

3.5 Similarity

We analyzed the diversity of information provided by each metadata type in the previous section. In this section, we study the pairwise relatedness of social bookmarks, anchor texts and search queries, i.e. how similar each metadata type is *to the others*. We also use categorization information from the Open Directory as ground truth to investigate whether each metadata type is suited for classification tasks, thereby extending the related studies in [4, 27, 28]. We used cosine similarity as similarity measure. The preprocessing of data involved splitting words at the special characters `, . : _ / # ! ? "`, followed by stemming based on Porter's stemming algorithm [23]. We also removed common English stop words such as "the" or "of" from the data. After these steps, each word was treated as one dimension in the vector space for similarity computation. The results are shown in table 3.

The highest similarities are between tags and categories (0.189) as well as between anchor text words and search query keywords (0.193)[9]. This direct comparison suggests

[8] Kendall-τ for tags, anchor words and search keywords are -0.91, -0.78 and -0.96, respectively.

[9] A statistical test reveals that the similarity means for (A, C) and (S, C)

	T	A	S	C
T	x	0.126	0.126	**0.189**
A	0.126	x	**0.193**	0.103
S	0.126	**0.193**	x	0.102
C	**0.189**	0.103	0.102	x

Table 3. *Similarity* Pairwise similarities of tags (T), anchor words (A), search keywords (S) and categories (C). The maximum values for each column are in bold font.

that tags are better suited for classification tasks whereas anchor words are better for augmenting web search.

Still, this does not mean that social annotations in general cannot improve web search. Au Yeung et al. [17] successfully used tags for web search disambiguation by exploiting the implicit semantics extracted from folksonomies. Additionally, we have shown in a previous work how to personalize web search by re-ranking search results lists based on the similarity of user and document profiles created from social annotations [20]. We therefore argue that social annotations can indeed help in the broad area of web search but it is important to verify whether they are the correct tool for solving a given problem.

3.6 Classification

We observed in the previous section that tags seem to be better suited for classification of web documents than anchor words or search keywords. In this section, we extend this analysis and study how each metadata type compares with the top-down categorization of the Open Directory.

We match tags, anchor words and search keywords of a document against its categorization. A document in the Open Directory is categorized by one or more category hierarchies such as "arts > crafts > textiles > weaving". We analyzed at which hierarchy depth matches occurred and normalized the results so that the top category in a hierarchy, e.g. "arts", is represented by 0 and the leaf category by 1, e.g. "weaving". Additionally, we used the Levenshtein distance [16] to relax the matching conditions in order to detect small variations such as singular-plural ("dog" vs. "dogs") or different languages ("music" vs. "músika") to a certain degree. The results are shown in figure 4.

Firstly, there are strong negative correlations with document popularity for all metadata types: Spearman-r for tags, anchor words and search keywords are -0.99, -0.84 and -0.99, respectively[10]. With increasing document popularity, broader classification scores are achieved. This also seems to indicate that popular websites cover rather broad topics whereas less popular websites are rather focused[11].

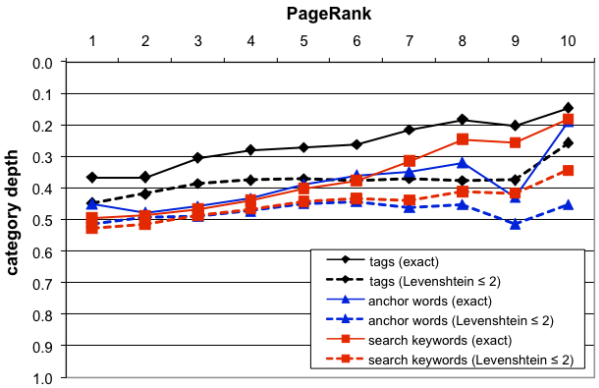

Figure 4. *Classification* Normalized category depth for matches of tags, anchor words and search keywords with categories. A value of 0 denotes a root category ("broad"), a value of 1 a leaf category ("specific"). The solid and dotted lines show exact matches and relaxed matches for a Levenshtein distance of up to 2, respectively.

Secondly, tags are used for broader categorization than anchor words and search keywords: the global average for matches of tags is 0.27 compared to 0.41 and 0.43 for anchor words and search keywords, respectively. Under relaxed matching conditions, tags score 0.38 compared to 0.47 for both anchor words and search keywords. This result supports the previous conclusion that tags are better suited for classification purposes than anchor words or search keywords in the sense that they can better catch the "aboutness" of documents (cf. [4, 10]). In the future, we plan to investigate how tags can help to identify multi-topic web documents.

4 Conclusion and Discussion

In this paper, we studied three different types of metadata about web documents: social annotations provided by readers of web documents, anchor text of incoming hyperlinks provided by authors of web documents, and search queries of users trying to find web documents. We created a large research data set from various web sources and used it to investigate several characteristics of said metadata including length, novelty, diversity, and similarity and discussed

are significantly different for $P<0.05$. For (A, T) and (S, T) however, the null hypothesis of having equal means could not be rejected.

[10] Kendall-τ for tags, anchor words and search keywords are -0.96, -0.73 and -0.96, respectively.

[11] For example, among the PR10 websites in our data set are WhiteHouse.gov and NASA.gov compared to websites such as WomencareShelter.org or LakeGeorgeRestaurants.com for PR3.

theoretic and practical implications. Scientists are invited to use our CABS120k08 data set for their own research.

While we consider this study as a good starting point, there are still opportunities for improvements and future work. For example, we did not yet include temporal information in our evaluations even though temporal data is included in the CABS120k08 data set. Additionally, our analysis of search queries is focused on information provided about web documents. Search query logs also give insights into user behavior and user preferences, an aspect which we did not include in this study.

On the other hand, our study has produced results that are worth further research. For example, we found that social bookmarking seems to be particularly helpful for identifying the "aboutness" of web documents. We plan to investigate in the future whether there is a relation of topical information derived from social bookmarking to the notion of "hubs" and "authorities" as described by the HITS algorithm for WWW link structure [14].

5 Related work

To the best of our knowledge, this is the first study analyzing social bookmarks/tagging, anchor texts and search queries in combination and direct comparison based on a large set of real-world data. Next to the references mentioned throughout the text, the following works are related to the work described in this paper.

Brin and Page described the first implementation of the Google search engine in 1998 [6]. They pointed out that anchor texts often provide more accurate descriptions of web documents than the documents themselves. We could measure in our study that anchor text does indeed provide meaningful information for information retrieval tasks but that it is less suited for catching the "aboutness" of web documents than data derived from social annotations. This suggests that the up-and-coming social annotations could prove to be a helpful addition to the information retrieval toolset.

Eiron and McCurley analyzed anchor text for web search based on a study of the IBM intranet [10]. They found that anchor text resembles real-world search queries with regard to term distribution and length. Our results confirm and measure the similarity of anchor texts and search queries. However, we found that anchor texts are generally less likely to be contained in a document's content: Eiron and McCurley reported 66.4% for full matches of anchor text with document content compared to 51.0% in our study.

Heymann et al. [12] studied whether social bookmarking can improve web search. Due to the different research focus their work includes only a short analysis of the novelty of tags with regard to document content or anchor text. While their results are difficult to compare with without more information and figures, our findings in this area seem to comply with theirs.

References

[1] D. Abrams, R. Baecker, and M. Chignell. Information archiving with bookmarks: personal web space construction and organization. In *Proceedings of SIGCHI '98*, pages 41–48, 1998.

[2] M. Ames and M. Naaman. Why we tag: motivations for annotation in mobile and online media. In *CHI '07: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 971–980, New York, NY, USA, 2007. ACM.

[3] R. Baeza-Yates and A. Tiberi. Extracting semantic relations from query logs. In *KDD '07: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 76–85. ACM, 2007.

[4] S. Bao, G. Xue, X. Wu, Y. Yu, B. Fei, and Z. Su. Optimizing web search using social annotations. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 501–510, New York, NY, USA, 2007. ACM Press.

[5] S. M. Beitzel, E. C. Jensen, D. D. Lewis, A. Chowdhury, and O. Frieder. Automatic classification of web queries using very large unlabeled query logs. *ACM Trans. Inf. Syst.*, 25(2):9, 2007.

[6] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *Proceedings of WWW7*, pages 107–117, Amsterdam, The Netherlands, The Netherlands, 1998. Elsevier Science Publishers B. V.

[7] S. Chakrabarti, B. Dom, and P. Indyk. Enhanced hypertext categorization using hyperlinks. In *SIGMOD '98: Proceedings of the 1998 ACM SIGMOD international conference on Management of data*, pages 307–318, New York, NY, USA, 1998. ACM.

[8] J. Dean and S. Ghemawat. Mapreduce: Simplified data processing on large clusters. *Sixth Symposium on Operating System Design and Implementation (OSDI)*, 2004.

[9] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6):391–407, 1990.

[10] N. Eiron and K. S. McCurley. Analysis of anchor text for web search. In *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in Information Retrieval*, pages 459–460, New York, NY, USA, 2003. ACM.

[11] S. A. Golder and B. A. Huberman. Usage patterns of collaborative tagging systems. *J. Inf. Sci.*, 32(2):198–208, 2006.

[12] P. Heymann, G. Koutrika, and H. Garcia-Molina. Can social bookmarking improve web search? In *Proceedings of 1st ACM International Conference on Web Search and Data Mining (WSDM'08)*, pages 195–206. ACM, February 2008.

[13] M.-Y. Kan. Web page categorization without the web page. In *WWW Alt. '04: Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters*, pages 262–263, 2004.

[14] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. In *SODA '98: Proceedings of the ninth annual ACM-SIAM symposium on Discrete algorithms*, pages 668–677, Philadelphia, PA, USA, 1998. Society for Industrial and Applied Mathematics.

[15] R. Kraft and J. Zien. Mining anchor text for query refinement. In *WWW '04: Proceedings of the 13th international conference on World Wide Web*, pages 666–674, New York, NY, USA, 2004. ACM.

[16] V. I. Levenshtein. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10(8):707–710, February 1966.

[17] C. man Au Yeung, N. Gibbins, and N. Shadbolt. Web search disambiguation by collaborative tagging. In *Proceedings of the Workshop on Exploring Semantic Annotations in Information Retrieval (ESAIR) at ECIR'08*, pages 48–61, 2008.

[18] C. Marlow, M. Naaman, D. Boyd, and M. Davis. Ht06, tagging paper, taxonomy, flickr, academic article, to read. In *Proceedings of HT '06*, pages 31–40, 2006.

[19] M. G. Noll and C. Meinel. Authors vs. readers: A comparative study of document metadata and content in the www. In *Proceedings of 7th Int'l ACM Symposium on Document Engineering '07*, pages 177–186, Winnipeg, Canada, 2007.

[20] M. G. Noll and C. Meinel. Web search personalization via social bookmarking and tagging. In *Proceedings of 6th International Semantic Web Conference and 2nd Asian Semantic Web Conference*, Busan, South Korea, 2007.

[21] M. G. Noll and C. Meinel. Exploring social annotations for web document classification. In *SAC '08: Proceedings of the 2008 ACM symposium on Applied computing*, pages 2315–2320, Fortaleza, Ceara, Brazil, 2008. ACM.

[22] G. Pass, A. Chowdhury, and C. Torgeson. A picture of search. In *Proceedings of 1st International Conference on Scalable Information Systems*, Hong Kong, 2006.

[23] M. F. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.

[24] A. Spink, B. J. Jansen, D. Wolfram, and T. Saracevic. From e-sex to e-commerce: Web search changes. *IEEE Computer*, 35(3):107–109, 2002.

[25] A. Spink, D. Wolfram, M. B. J. Jansen, and T. Saracevic. Searching the web: The public and their queries. *Journal of the American Society for Information Science and Technology*, 52(3):226–234, 2001.

[26] B. Wu and B. D. Davison. Identifying link farm spam pages. *Special interest tracks and posters of the 14th international conference on World Wide Web (IW3C2)*, 2005.

[27] X. Wu, L. Zhang, and Y. Yu. Exploring social annotations for the semantic web. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pages 417–426, New York, NY, USA, 2006. ACM Press.

[28] S. Xu, S. Bao, Y. Cao, and Y. Yu. Using social annotations to improve language model for information retrieval. In *CIKM '07: Proceedings of the 16th ACM conference on Conference on information and knowledge management*, pages 1003–1006, New York, NY, USA, 2007. ACM.

[29] Z. Zhuang and S. Cucerzan. Re-ranking search results using query logs. In *CIKM '06: Proceedings of the 15th ACM international conference on Information and knowledge management*, pages 860–861, New York, NY, USA, 2006. ACM.