Alternative Searching Services:

Seven Theses on the Importance of "Social Bookmarking"

Dr. Gernot Graefe

Business Development, Corporate
Computing and Communication Lab
Fuerstenallee 11
33102 Paderborn
gernot.graefe@c-lab.de

Dr. Christian Maaß
Dr. Andreas Heß

Lycos Europe
Carl-Bertelsmann-Str. 29
33311 Guetersloh {christian.maass,
andreas.hess}@lycos-europe.com

Abstract: In recent years social bookmark systems like del.icio.us or Furl have become increasingly popular. These systems sometimes are regarded as alternatives to algorithmic search engines like Google. In this paper we develop seven theses on the potential of these systems in order to establish a conceptual basis for future research in this area. Thereby it becomes clear that social bookmarking systems complement rather than threaten algorithmic search engines.

1 Introduction

Together with the exponential growth of the Internet, algorithm-based search engines such as Google or Microsoft Live Search have become the most frequented Web applications. According to conservative estimates 70 to 85% of all information inquiries are serviced by such search engines [HD04]. Their success is mainly based upon their ability to index information automatically and provide it to a great number of users independent of time and place. However, there is empirical evidence that the quality of their search results is rather low. Frequently, only 20 to 45% of the search engine results are relevant results considering the supplied information inquiry [MW03]. One explanation why only such a small portion of the algorithm-based search engine's results are relevant hits are search engine manipulations. As an example one may mention BMW, which after an all too obvious attempt to manipulate Google's search engine algorithms had been temporarily banned from Google's index in early 2006 [BBC06].

Against this background it does not come as a surprise that more and more people wonder whether alternative searching services can compete with algorithm-based search engines with regards to quality of search results [Ne05]. In the media bookmarking systems are already considered as alternatives to Google and other algorithm-based search engines [MNBD06].

However, it is surprising that there are only very few studies on search engine quality [LH07]. This paper shall therefore try to assess the potential of social bookmarking systems. To this end it has to be determined whether social bookmarking systems are in fact alternatives to algorithm-based search engines, and which weaknesses and strengths they possess compared to algorithm-based search engines. First this paper will outline the way social bookmarking systems work and then go on to develop methods and criteria to evaluate the quality of search engines' information retrieval. Finally, seven theses on the importance of social bookmarking systems will be elaborated in order to establish a conceptual basis for future research in this area. The paper ends with a short conclusion.

2 Technological Foundation and Methodical Background

2.1 Characterization of Algorithm-Based Search Engines and Social Bookmarking

In order to access the quality of different searching services we shall first outline the way they work as well as their differences in information retrieval and analysis. Algorithm-based search engines make use of technological resources. So-called Web crawlers automatically analyze the World Wide Web by autonomously following all hyperlinks that are placed on a particular Web page. This allows them to analyze a great part of the Internet and index it for later search inquiries in a rather short period of time. The hyperlinks and page information that the robots can gather are then saved in a special database, the so-called index. This index and the enormous amount of stored data are then used to generate search results for every search inquiry.

Compared to these search engines, social bookmarking systems – such as del.icio.us or Furl – are quite new, and are only discussed in the general public recently. Hence it not surprising that no commonly accepted definition for this term has been established. In principle one may point out that social bookmarking systems are a special form of social software solutions that are used to create social networks and distribute information within these networks [ANRD06].

These social networks play an important role in the context of social bookmark systems as the index is not build by a Web crawler but through the collaboration of the network's members. For this purpose members only need to publish their personal hyperlink collections, or a fraction of it, in the respective bookmark system. Furthermore each hyperlink should be "tagged" with metadata that serves as a description of the particular Link and corresponding Web site [SLRC06]. For example a hyperlink to the "White House" in Washington could be published with the tags "President", "USA", "White House", "sightseeing in Washington" and "George Bush". Through these tags, a search inquiry for the President of the United States could illustrate a connection between the President and the White House, even if the respective Web-document does not reveal such a connection. The technology behind this is based on an analysis of the relation between the tags that reveals the frequency of their joint usage.

The emerging networks of relationships between tags and hyperlinks are called folksonomies. They enable the user to navigate through a collaboratively elaborated index. Figure 1 shows the principles of social networking systems on the basis of the above text. Evidently, these systems are quite different from traditional bookmarking systems.

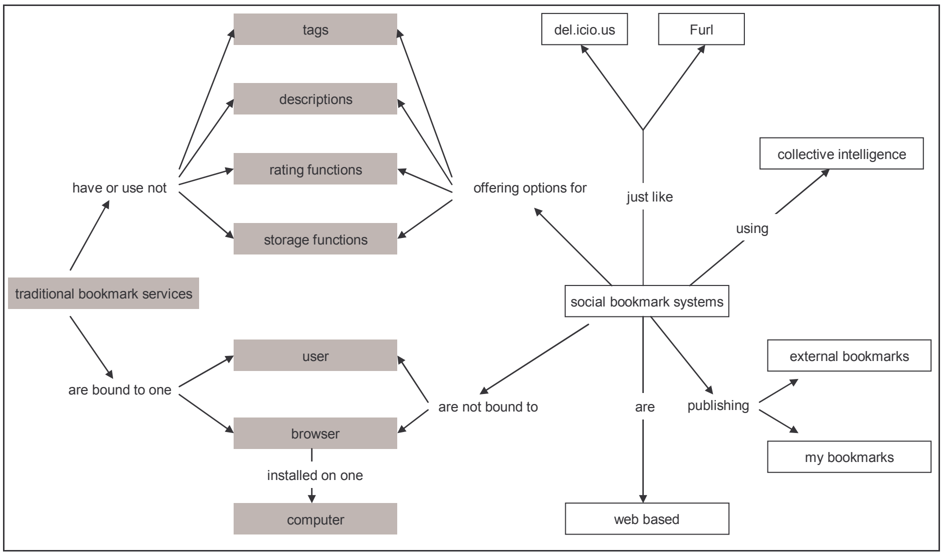

Figure 1: Comparison between traditional and social bookmarking systems [translated from Dö07]

2.2 Current Research on Web-Information Retrieval

Information retrieval constitutes a special area of computer science which deals with the computer-based and content-oriented extraction of information. In order to pay special attention to the peculiarities of retrieving information from the Web and due to the increasing significance of this work, Web-information retrieval now forms an own research area. The assessment of search engine quality is an important part of this research that is usually measured through so called retrieval tests. In these tests search inquiries are sent to selected search engines. Afterwards a panel of experts evaluates the supplied search results according to their relevance [Gr04; Ve06] whereby the so called Precision has become an accepted and commonly used measuring unit. This index number points out the percentage of relevant hits within the total number of search results. However, the mere focus on this percentage is criticized to an increasing degree. It is argued that other measures have to be taken into account in order to gain valuable information on the quality of search engine results [LH07]. Of particular significance are the size and the up-to-datedness of the search index that determines which information the user has access to.

One can conclude from the above text that it is primarily the size and actuality of the search index as well as the search results' relevance that matter in the assessment of search engine quality. Thus these criteria will be used in order to compare algorithm-based search engines and social bookmarking systems in the remainder of this paper. Thereby seven hypotheses will be developed that point out the strength and weaknesses of social bookmarking systems and set a conceptual framework for further empirical studies.

3 The Quality of Social and Algorithm-Based Searching Services

3.1 Size of the Index and the Importance of It Being Up-to-Date

The size of the searching index and its up-to-datedness have already been identified as important indicators of searching service quality. Considering that the indexed part of the World Wide Web currently includes over 11.5 billion Web sites [GS05], it seems arguable that the manual compilation of Web site information done by a community in a social bookmark system could somehow be better than that of automated algorithm-based search engines. This assumption can be strengthened through a direct comparison of the two respective services' search index sizes. Empirical studies show that the indices of algorithm-based search engines of Google, Yahoo and MSN covering nearly 85% of the part of the Internet that can be indexed [Su06]. Hence, several billion Web sites are indexed. In contrast, the leading social bookmarking system in Germany, Mister-Wong, had only indexed about 1.4 million Web sites in early 2007 [Mi07]. Moreover, towards the topics covered, social bookmark systems seem to have a technical and media-oriented focus right now that limits their coverage. Other topics of interest are hardly covered. The existence of such a limited focus can be deduced from an analysis of the most frequently used tags and the respective number of bookmarks in these areas Figure 2 shows an analysis of the bookmark system provided by Lycos-Europe.

Thesis 1: Social bookmarking systems currently cover only a limited number of subjects/ topics on the Internet. However, as social bookmarking systems grow they will continuously widen their range of subjects.

It is not enough, however, to draw one's conclusion on the quality of searching services merely on the basis of their index size as search engine robots index all available Web sites. In contrast social bookmarking systems disregard poor Web sites in a pre-selection. Consequently, the indices of social bookmarking systems will always be smaller simply due to the way they work. Just because of this filter function social bookmarking systems may provide result lists that have a higher Precision with respect to the initial inquiry.

Thesis 2: The smaller index of social bookmarking systems does not correlate with the perceived quality of the respective searching services.

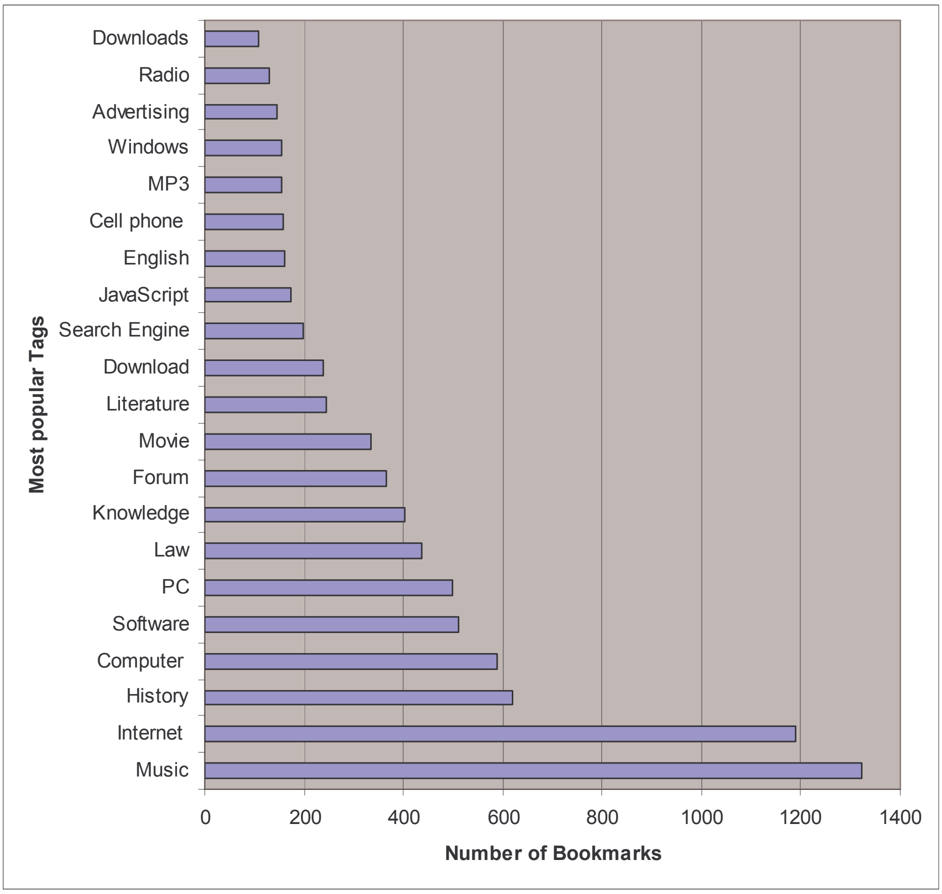

Figure 2: Most frequently used tags in the social bookmarking system Lycos IQ [Ly07]

Aside from the index size, the up-to-datedness of the index is commonly identified as a yardstick for searching service quality. That is because new information is the most looked for information on the Internet, just like business news, sports articles or job opportunities. The value of this information strongly correlates with its up-to-datedness. Thus it seems as a logical approach to consider the frequency in which searching services are updated next.

The average frequency with which search engine providers update their indices is about once every 3.1 days for Google, once every 3.5 days for MSN, and Yahoo updates its index every 9.8 days [LWM06]. Some smaller search engines only update their index in intervals greater than 30 days, which seems unacceptable considering the speed with which content changes on the Internet: 320 million new Web sites are being created per week and within a year about 80% of all Web sites change their link-structure [NCO04].

However, to some extend the high average intervals in which search engine providers update their indices can be attributed to the fact that the providers maintain separate data storages for specific areas of interest [NCO04]. For example, the news index of search engines such as Google and MSN is updated daily. In contrast, the index for image searches is updated in much greater intervals. Though, algorithm-based search engines have a planned renewal strategy for their indices. It is arguable whether social bookmarking systems can deliver recent news to the extent that search engines do. For the integration of current information it is necessary that users add links to their list of bookmarks. As for the short lifespan of a specific piece of current information one may have doubts if users will actually bookmark a specific notice referring to this specific information. It seems much more likely that users will add Web sites to their bookmarks that regularly provide current news. For example the Web site of a news agency will be bookmarked rather than an individual notice on this Web site. In this case it would be impossible to find a dedicated notice by the means of social bookmarking systems.

Thesis 3: Algorithm-based search engines are able to include information into their indices faster and more detailed than social bookmarking systems.

Thesis 4: The greater and more active a community is, the more likely it is to find sites that contain a dedicated notice through search results of social bookmarking systems.

3.2 Relevance of Search Results

Algorithm-based search engines access the relevance of Web sites primarily based upon two factors. Firstly, the analysis of the different elements inside the HTML-code plays an important role in determining the sites' relevance [GC06]. Search engine robots weight the respective elements differently, meaning that not every element has the same impact on the Web site's search engine rating. Take for example keywords that are used in headlines. Based upon the assumption that these keywords summarize the site's content precisely, they are well suited for an assessment of the Web page's contextual relevance. Consequently, text that is declared as a headline in the HTML-code weights heavier in the site's relevance assessment than conventional text does. However, the page content is also very important for the determination of the site's relevance. For this reason many companies try to place frequently used search terms on their Web page. If the search engine locates the respective search term very often, the corresponding Web page will be ranked high in the search engine result list for this particular keyword. Yet the ranking based upon this criterion alone has been subject to several manipulation attempts as popular keywords were systematically integrated into the HTML-code of the Web pages in order for them to receive better ratings.

Secondly, nearly all search engines analyze the link-structure in order to evaluate the contextual relevance and quality of Web pages. It is believed that sites with popular or high quality content receive a higher number of hyperlinks as compared to Web sites with inferior content. In combination with contextual criteria – such as keywords – the link-structure is able to significantly improve the quality of search results.

A few years ago, Google was able to conquer the search engine market due to their implementation of this groundbreaking idea [BP98]. Nevertheless, also this criterion is not resistant to manipulations. Cloaking is one attempted to manipulate the ratings. Cloaking means that special software solutions on Web servers try to distinguish human users from search engine robots. The latter are than forwarded to a special search engine optimized Web page with hyperlinks and keywords that tricks the robot into assuming that it has found a highly relevant page. Without wanting to start a detailed discussion on the problem areas of algorithm-based search engines, we may conclude from the above discussion that algorithm-based search engines depend on criteria that are vulnerable to manipulation attempts. In part, this explains the low Precision of their search results.

Unlike search engines, social bookmarking systems are less vulnerable to manipulation attempts. Here, the contextual relevance of Web pages is not accessed through robots but humans. For the users, it is not the HTML-code elements or the link-structure that affects them to add a Web page to their personal bookmark list. Rather it is the information quality of the respective Web page. For this reason social bookmarking systems base their ratings upon the cumulative number of users that have bookmarked a certain Web page. Therefore Web sites may be ranked high even if there are very few links that lead to this document. In principle, these bookmarking systems are comparable with customer reviews on shopping portals like Amazon that are already used for a fairly long time. This is an important note as these customer reviews are regarded as particularly trustworthy [Eg01; Ni02]. Social bookmarking systems use this factor in order to increase the trustworthiness of their search results as the community disregards low quality content.

Thesis 5: Compared to algorithm-based search engines, social bookmarking systems are far less prone to manipulations. This results in a greater Precision of search inquires.

Thesis 6: Users perceive the search results of social bookmarking systems as more trustworthy than those of algorithm-based search engines.

Still, this rather favorable assessment of social bookmarking systems is put into perspective by the fact that problems in terms of “tagging” are quite frequent. For example, in the course of an analysis of the leading social bookmarking system delicio.us, Lee points out that about 20% of its users do not annotate/ tag any of their bookmarks [Le06]. Moreover, different spellings and subjective combinations of tags lead to more or less diffuse folksonomies. Therefore frequently errors occur while searching for connected issues and subjects. However, it may be assumed that problems originating due to different spellings may be solved by technical means in the near future. For example, algorithm-based search engines are comparing search inquiries with a predetermined vocabulary, immediately identify spelling mistakes and instantly suggest an orthographically correct word to the user. Furthermore several research projects try to address the above mentioned problems by connecting semantic technologies with social software solutions. It is one aim of these projects to automatically extract a Web sites metadata to facilitate the tagging for the user [WZY06].

Thesis 7: The quality of tags will be improved in the near future through semantic technologies.

3 Conclusion

Following the preceding discussion, the strength of social bookmarking systems can be seen in their ability to evaluate the quality of Web sites better than algorithm-based search engines. In addition to that, context relevant connections can be created through metadata/ tags that annotate links and Web pages. While algorithm-based search engines are unable to determine the amplitude and correctness of information encountered on a Web site, the users of social bookmark communities can. They may evaluate a Web site and then share this evaluation with other members of the community. For future retrieval test this means that the currently used technical measures have to be extended by other measures that are able to deliver a better evaluation of the information quality and focusing on the content itself. Nevertheless there are still doubts whether or not social bookmarking systems can compete with algorithm-based search engines with regard to the indexation of current information. It is questionable if users will bookmark Web sites that contain information with a short lifespan. To overcome this problem the automatic generation of metadata could be a visible approach. As shown by Hess et al. such an approach could be a first step to improve and enhance the manual approach of meta data generation in terms of quantity and quality [HMD07].

Upon this background one may resume that social bookmarking systems do not replace algorithm-based search engines. Rather they can be treated as qualitative complement of traditional searching services. Therefore it seems to be a feasible approach for algorithm based search engine providers to integrate the results of social bookmarking systems into their search process, in order to improve the quality of their search results. This approach is already used by search engines like Lycos. However, until now there is no empirical data if such an approach could improve the quality of search results. Furthermore it would be an interesting approach to use the folksonomies generated by the community as a starting point towards the realization of structured vocabulary just like in the field of ontology engineering. As the classic top-down ontology-based approach has not been widely adopted due to its complexity in real-world use, the public has clearly indicated a strong preference for bottom-up approaches using loose folksonomies instead.

4 Acknowledgements

The research presented in this paper was funded by the German Federal Ministry of Economy and Technology (BMWi) under the THESEUS project.

Bibliography

[ANRD06] Aleman-Meza, B.; Nagarajan, M.; Ramakrishnan, C.; Ding, L.; Kolari, P.; Sheth, A.; Arpinar, B.; Joshi, A.; Finin, T: Semantic Analytics on Social Networks: Experiences in Addressing the Problem of Conflict of Interest Detection. In: Proceedings of the 15th International Conference on World Wide Web, Edinburgh, May 23 - 26, 2006. ACM Press, New York, pp. 407-416.

[BBC06] BBC News: BMW given Google 'death penalty', 2006; published on the Internet: http://news.bbc.co.uk/2/hi/technology/4685750.stm (accessed April 9th, 2007).

[BP98] Brin, S.; Page, L.: The Anatomy of a Large-Scale Hypertextual Web Search Engine, Stanford University, 1998.

[Dö07] Döbeli, B.: Social Bookmarking, 2007; published on the Internet: http://beat.doebe.li/bibliothek/w01899.html (accessed April 9th, 2007).

[Eg01] Egger, F. N.: Affective Design of E-Commerce User Interfaces: How to Maximise Perceived Trustworthiness. In: Helander, M. G.; Khalid, H. M.; Tham, M. P. (Eds.): Proceedings of the International Conference on Affective Human Factors Design. Asean Academic Press, London, 2001.

[GC06] Grappone, J.; Couzin, G.: Search Engine Optimization, Sybex, 2006.

[Gr04] Griesbaum. J.: Evaluation of three German search engines: Altavista.de, Google.de and Lycos.de. In: Information Research 9, 4, 2004; published on the Internet: http://informationr.net/ir/9-4/paper189.html (accessed April 9th, 2007).

[GS05] Gulli, A.; Signorini, A.: The Indexable Web is more than 11.5 Billion Pages. In: Special Interest Tracks and Posters of the 14th International Conference on World Wide Web, Chiba, 2005.

[HD04] Hirsh, S.; Dinkelacker, J.: Seeking information in order to produce information: An empirical study at Hewlett Packard Labs. In: Journal of the American Society for Information Science and Technology 55, 9, 2004; pp. 807-817.

[HMD07]: Heß, Andreas; Maaß, Christian; Dierick, Francis (2007): On Semi-Automated Semantic Tagging of Very Short Texts, Lycos Research Paper 1/2007, Gütersloh 2007.

[Le06] Lee, K.: What Goes Around Comes Around: An analysis of del.icio.us as social space. In: Proceedings of the 20th anniversary conference on Computer supported cooperative work, 2006; published on the Internet: http://delivery.acm.org/10.1145/1190000/1180905/p191-lee.pdf?key1=1180905

&key2=7444766711&coll=&dl=ACM&CFID=15151515&CFTOKEN=6184618 (accessed April 9th, 2007).

[LH07] Lewandowski, D.; Höchstötter, N.: Web Searching: A Quality Measurement Perspective. In: Spink, A.; Zimmer, M. (eds.): Web Searching: Interdisciplinary Perspectives. Springer, Dordrecht 2007.

[LWM06] Lewandowski, D.; Wahlig, H.; Meyer-Bautor, G.: The Freshness of Web search engine databases. In: Journal of Information Science 32, 2, 2006; pp. 131-148.

[Ly07] Lycos: Link library, 2007; published on the Internet: http://iq.lycos.de/lili/srch/ (accessed April 9th, 2007).

[Mi07] Mister-Wong: Mister-Wong – Startseite, 2007; published on the Internet: http://www.mister-wong.de/ (accessed April 9th, 2007).

[MNBD06] Marlow, C.; Naaman, N.; Boyd, D.; Davis, M.: HT06, Tagging Paper, Taxonomy, Flickr, Academic Article. To Read. In: Proceedings of the seventeenth conference on hypertext and hypermedia, 2006; pp. 31-40.

[MW03] Machill, M.; Welp, Carsten: Wegweiser im Netz, Bertelsmann Stiftung, Gütersloh, 2003.

[NCO04] 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 the 13th WWW Conference, New York, 2004; published on the Internet: www2004.org/proceedings/docs/1p1.pdf (accessed April 9th, 2007).

[Ne05] Neymanns, H.: Suchmaschinen: Das Tor zum Netz, Bundestagsfraktion der Grünen, Berlin, 2005; published on the Internet: http://www.gruene-bundestag.de/cms/publikationen/dokbin/63/63265.pdf (accessed April 9th, 2007).

[Ni02] Nikander, P.: Trustworthiness as an Asset. In: Schubert, S.; Reusch, B.; Jesse, N. (Hrsg.): Informatik bewegt, Tagungsband der 32. Jahrestagung der Gesellschaft für Informatik e.V., 2002; pp. 100-105.

[SLRC06] Sen, S.; Lam, S.; Rashid, A.; Cosley, D.; Frankowski, D.; Osterhouse, J.; Harper, F.; Riedl, J.: Tagging, communities, vocabulary, evolution. In: Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work, Alberta, 2006; pp. 181-190.

[Su06] Sullivan, D.: Nielsen NetRatings Search Engine Ratings, Search Engine Watch, 2006; published on the Internet: http://searchenginewatch.com/showPage.html?page=2156451 (accessed April 9th, 2007).

[Ve06] Véronis, J.: A comparative study of six search engines, 2006; published on the Internet: http://www.up.univ-mrs.fr/veronis/pdf/2006-comparative-study.pdf (accessed April 9th, 2007).

[WZY06] Wu, X.; Zhang, L.; Yu, Y.: Exploring social annotations for the semantic web. In: Proceedings of the 15th International Conference on World Wide Web, New York, 2006; pp. 417-426.