Towards Improving Web Search by Utilizing Social Bookmarks

Yusuke Yanbe, Adam Jatowt, Satoshi Nakamura, and Katsumi Tanaka

Department of Social Informatics, Kyoto University
Yoshida-Honmachi, Sakyo-ku, 606-8501
Kyoto, Japan
{yanbe,adam,nakamura,tanaka}@dl.kuis.kyoto-u.ac.jp

Abstract. Social bookmarking services have become recently popular in the Web. Along with the rapid increase in the amount of social bookmarks, future applications could leverage this data for enhancing search in the Web. This paper investigates the possibility and potential benefits of a hybrid page ranking approach that would combine the ranking criteria of PageRank with the one based on social bookmarks in order to improve the search in the Web. We demonstrate and discuss the results of analytical study made in order to compare both popularity estimates. In addition, we propose a simple hybrid search method that combines both ranking metrics and we show some preliminary experiments using this approach. We hope that this study will shed new light on the character of data in social bookmarking systems and foster development of new, effective search applications for the Web.

Keywords: Web search, social bookmarks, PageRank, meta-search.

1 Introduction

In the early years of the Web, directory services were utilized in order to arrange the Web and to make it accessible to users. However, the rapid growth of the Web soon made this approach impractical. Computing page relevance was also insufficient since usually too many pages were relevant to user queries. In order to effectively rank pages the quality of Web documents had to be captured. Thus, came the era of link based algorithms such as PageRank [18] and HITS [10], which estimate quality of pages by measuring their relative popularity in the Web. PageRank is currently the most popular link-based Web page ranking method. It is based on a random surfer model, where the probability of the surfer reaching a given page is calculated as the result of a random selection of links. Consequently, the popularity of the page is determined on the basis of the size of its hypothetical user stream.

Link-based page popularity estimation has, however, several disadvantages. One is related to the difficulty of creating links as it usually requires some effort and knowledge from users. Although, recently we observe the explosion of Weblogs or wikis, which make the link creation relatively easy, yet search engines seem not to trust links on such pages due to spamming threats. In general, links on majority of

L. Baresi, P. Fraternali, and G.-J. Houben (Eds.): ICWE 2007, LNCS 4607, pp. 343–357, 2007.

pages are still created by a relatively small group of content producers. However, there is an overwhelming group of content consumers whose opinions cannot be captured by standard link-based ranking metrics. Additionally, links often serve many different purposes on Web pages and, hence, should not always be treated as positive votes for pages [14].

Another disadvantage of link based ranking mechanisms is related to their temporal aspect. Web is a very dynamic environment and many new pages are continuously created (see for example, [7,17]). However, pages usually need long time to acquire links and to become popular among Web authors. In result, PageRank algorithm is biased against new pages [2,12,23]. Considering the fact that users usually need fresh information, this bias is a major disadvantage of link-based algorithms making them weak in providing fresh content or detecting new, hot topics and trends in the Web.

In general, the conventional link-based ranking approach is still useful, mostly due to its success in combating spamming, however, we believe that it needs to be complemented by another reliable metric. Along with the advent of Web 2.0, social bookmarking systems seem currently to have a potential for improving the capabilities of existing search engines. Social bookmarking lets users share, classify, and discover interesting Web pages. In social bookmarking systems, the popularity of a Web page is usually calculated by the total number of times it has been bookmarked by users. We call this measure SBRank. As creating bookmarks is relatively easy and does not require much technological knowledge, thus, in contrast to links, any Web user can freely vote for pages. This, together with the high level of social interaction in social bookmarking services, makes SBRank a highly dynamic measure allowing for detecting high-quality, fresh and hot information on the Web.

Although social bookmarks have many advantages, relying on them alone is still not advisable in a general purpose Web search. This is because there is currently not enough data to produce satisfactory results for any arbitrary query[1]. Although, recently, we are observing a rapid increase in the number of bookmarked pages, yet we believe that the combination of link structure and social bookmarking based popularity estimates seems to be currently an optimal strategy. Future search applications should have at least the scalability of the existing search engines combined with improved ranking models.

In this paper, we attempt to make a comparative analysis between PageRank and SBRank metrics. The objective of this investigation is to analyze the feasibility and potential of a hybrid search method that would combine both popularity measures. In order to do so, we examine pages in social bookmarking systems and analyze their popularity using SBRank and PageRank measures. We also investigate the dynamics of SBRank metric in order to analyze whether it can improve freshness of search results.

More socially-aware search algorithms that would leverage the content of so-called Web 2.0 are an attractive vision as users often want to find information that is socially accepted (recommended by many users) and also recently popular. However, conventional link-based ranking methods cannot completely fulfill such requirements.

[1] Meta-search applications that would increase the amount of data by collecting evidences from different social bookmarking services have not appeared yet.

This work attempts at laying foundations towards building Web search applications that would exploit social bookmarks. We believe that our analysis and other similar systematic studies are necessary for designing reliable and high-quality Web search applications.

Previous studies of social bookmarking in the Web focused mostly on its social and linguistic aspects [6,15,16,19,20,21,22]. For example, the phenomenon of folksonomy (i.e. community-evolved taxonomy) was analyzed [16,19,21,22], tagging dynamics was examined [6] or a taxonomy of the current social bookmarking services was proposed [15]. The aim of our investigation is, however, different from these works, and has a practical objective, that is, examining the possibility and potential benefits of complementing traditional Web search with social bookmarking data.

The rest of this paper is organized as follows. Section 2 discusses the related research. Section 3 demonstrates the results of the analysis that we made. Next, Section 4 summarizes our findings and discusses the issues involved with building Web search applications that would utilize social bookmarks. Lastly, Section 5 concludes the paper and provides a brief look at our future work.

2 Related Work

The origins of social bookmarking date back to the work of Keller et al. [9] who in 1997 proposed to enhance Web browsers' bookmarking capabilities by using collaborative approach. Later, Bry and Wagner [3] also conducted a similar research. In the end of 2003, Joshua Schachter launched the first social bookmarking service called del.icio.us[2]. Later, many kinds of social bookmarking systems have been established and, currently, we are witnessing a rapid increase in their popularity.

Although, already some investigations have been made [6,15,16,19,20,21,22], social bookmarking is still a relatively new phenomenon that has not been studied well. Studies that have been made so far focused mostly on the issues related to folksonomy and social aspects. For example, Zhang et al. [22] introduced a hierarchical concept model of folksonomies using HACM - a hierarchy-clustering model. The authors reported that certain kinds of hierarchical and conceptual relations exist between tags. In another work, Golder and Huberman [6] measured regularities in user activities, tag frequencies, and bursts in popularity of tags used in social bookmarks. The authors discussed also dynamics of tagging exhibited in social bookmarking. In addition, tags were classified into seven categories depending on the functions they perform for bookmarks. More recently, Marlow et al. [15] introduced the taxonomy of tagging systems to illustrate their potential benefits. In another work, Wu et al. proposed a search model for annotated Web resources using social bookmarks as an example [20]. Nevertheless, none of the previous studies made comparative analysis of link- and social bookmark-based page ranking methods for the purpose of their combination.

Recently, several researches have been done on temporal link analysis [1,2,4]. Temporal link analysis focuses on link evolution, discovering link change patterns or on utilizing link timestamps for improving page ranking. For example, Amitay et al.

[2] http://del.icio.us/

proposed a method for finding authority pages in time as well as for detecting trends in the Web by using link timestamps [1]. Baeza-Yates et al. [2] suggested modifying PageRank by incorporating last-modification dates of pages. The objective was to eliminate the bias of PageRank towards old pages [2,12,23]. In another paper, Cho et al. [4] proposed a quality model of pages based on the changes in the amount of in-bound links of pages in time. According to this model, pages with growing popularity trends, measured by large increases of in-bound link numbers, should have highest qualities assigned, especially, if they are still relatively unpopular in the Web. On the other hand, Yu et al [23] proposed an algorithm called Timed PageRank for incorporating link duration into page ranking process by exponentially decaying PageRank scores of linking pages. However, the approaches that use links dynamics are rather impractical as it is usually difficult to determine link creation dates. In contrast, social bookmarks usually contain timestamps indicating dates of their creation. Thus, unlike in the case of link-based ranking, incorporating temporal aspects into the Web search seems to be generally more feasible by using social bookmarks.

Lastly, meta-search engines [5,11,13] are also related to our work. Several meta-search engines have been recently employed on the Web. They provide the advantage of the increased coverage of the Web as well as more up-to-date results due to drawing data from multiple search engines. No approach has been, however, proposed so far to combine the information derived from link structure and social bookmarks for enabling a joint page ranking metric. This is probably due to different characteristics of both information sources and the lack of their comparative analysis. In this paper, we attempt to fill in this gap.

3 Comparative Analysis

3.1 Dataset Characteristics

To analyze characteristics of pages in social bookmarking services we collected two datasets. As a source of the first dataset we selected del.icio.us since it is currently the most popular social bookmarking service[3] and it was also used by other researchers for studying social bookmarking [6,22]. Second dataset was created using Hatena Bookmark[4] – the most popular bookmarking service[5] in Japan, which was available online since February 2005.

Both datasets were obtained in the following way. We have utilized *popular tags*, which are sets of the most popular and recently used tags. Such tags are continuously published by del.icio.us[6] and Hatena Bookmark[7]. In total, *140* tags were retrieved on December 6th, 2006 from del.icio.us and *742* tags on February 16th, 2007 from Hatena Bookmark. Next, we collected popular URLs from these tags. Usually less

[3] In September 2006 it was reported that the service had 1 million registered users: http://blog.del.icio.us/blog/2006/09/million.html
[4] http://b.hatena.ne.jp
[5] The service had 60,000 users in October 2006: http://d.hatena.ne.jp/naoya/20061020
[6] http://del.icio.us/tag
[7] http://b.hatena.ne.jp/t

than *25* popular pages were listed for each tag in both social bookmarking systems. At this stage, we obtained *2,673* pages for del.icio.us and *18,377* pages for Hatena Bookmark. In the last step, we removed duplicate URLs (i.e. URLs listed under several popular tags). Finally, we obtained *1,290* and *8,029* unique URLs for del.icio.us and for Hatena Bookmark, respectively. Each URL had two attributes: firstDate and SBRank. firstDate indicates the time point when a page was introduced to the social bookmarking system for the first time by being bookmarked by one of its users. SBRank, as mentioned above, is the number of bookmarks of a given page obtained at the date of the dataset creation.

In order to detect PageRank values of the URLs, we used Google Toolbar[8] which is a browser toolbar that allows viewing PageRank values of visited pages[9]. PageRank values obtained in this way are approximated on the scale from *0* to *10* (*0* means the lowest PageRank value of a page).

To sum up, the obtained datasets are snapshots of the collections of popular pages in both social bookmarking systems. Each page has its Pagerank and SBRank values recorded which it had at the time of the dataset creation.

3.2 Distribution of PageRank and SBRank

Figures 1a and 1b show the percentage distribution of PageRank values in both datasets. We found that more than a half of pages (*56.1%*) have PageRank values equal to *0* in the del.icio.us dataset. There are even more such pages in Hatena Bookmark dataset (*81%*); probably due to its more local scope. These pages are rather unpopular according to the link-based ranking and are relatively difficult to be found using conventional search engines. However, many social bookmarkers considered them to be of high quality and bookmarked them in the systems. It may imply that the pages were discovered by users from other sources than conventional Web search engines. Possibly this could happen by interacting with the social bookmarking systems, since unlike bookmarks on a personal Web browser, social bookmarks affect users socially. For example, del.icio.us informs users about popular pages that recently obtained relatively many bookmarks[10]. Users can also subscribe to "Inbox" - a bookmark activity reporting service. From this feedback, pages attracting much attention can become rapidly known to many users.

In general, we think that there may be two possible reasons that caused the occurrence of many pages with low PageRank values in the datasets, despite their high popularity among social bookmarkers.

- The pages were created recently, thus, on average, they have relatively few inbound links.
- The pages were created long time ago but their quality cannot be reliably estimated using PageRank measure.

In order to determine which of these two reasons is more probable we did temporal analysis of the pages, which we will discuss in Section 3.4.

---

[8] http://toolbar.google.com

[9] We had to use Google Toolbar since Google API does not provide any automatic method for acquiring PageRank scores.

[10] http://del.icio.us/popular/

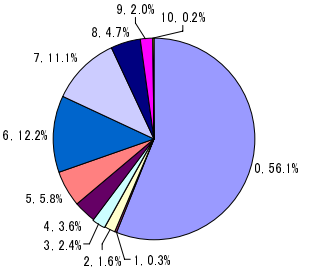

Fig. 1a. Distribution of PageRank values (del.icio.us)

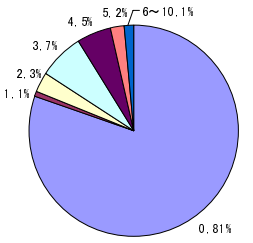

Fig. 1b. Distribution of PageRank values (Hatena Bookmark)

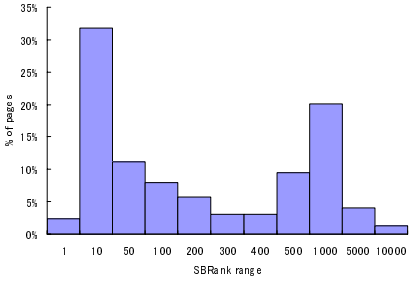

Fig. 2a. Histogram of SBRank (del.icio.us)

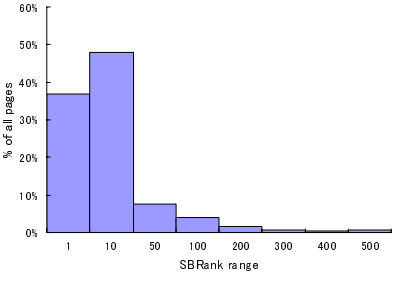

Fig. 2b. Histogram of SBRank (Hatena Bookmark)

Figures 2a and 2b show the distributions of SBRank values. It can be seen that quite few pages are bookmarked by many users, while the rest is bookmarked by a relatively low number of users. This is similar to PageRank metric that features power law distribution of PageRank values.

3.3 Correlation Between PageRank and SBRank

In this section we examine whether there is any correlation between PageRank and SBRank values. Figures 3a and 3b show scatter plots of both measures.

We observed a positive correlation coefficient ($r=0.53$ in del.icio.us and $r=0.10$ in Hatena Bookmark datasets) between SBRank and PageRank values. This is an important result, since, if the correlation coefficient had a very high value, that is, if generally SBRank values followed PageRank values, it would mean that PageRank alone adequately measures page quality. Hence, there would be no reason for its complementation with SBRank. On the other hand, if correlation coefficient between both measures had a very low absolute value, it would suggest that one of the metrics likely provides incorrect results. Since the values of the correlation coefficient were within the acceptable range, we can consider a combination of both rank estimates to be possible.

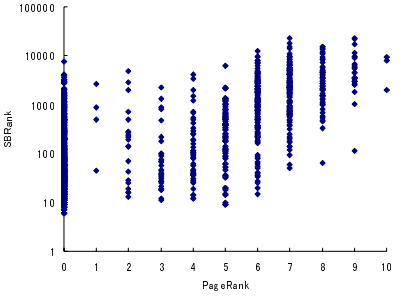

Fig. 3a. Scatter plot of PageRank and SBRank (del.icio.us) (logarithmic scale)

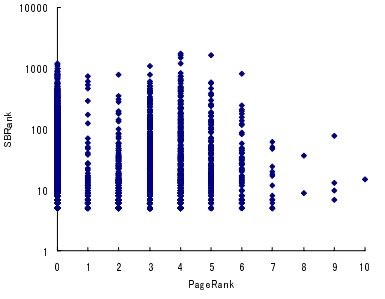

Fig. 3b. Scatter plot of PageRank and SBRank (Hatena Bookmark) (logarithmic scale)

3.4 Temporal Analysis

We turn now our attention to temporal aspects of the datasets. Figures 4a and 4b show plots of pages against the dates of their addition to the social bookmarking systems (firstDate). To correctly interpret these figures we have to remember that the datasets contain pages which were popular in both social bookmarking systems at the dates of the datasets' creation (December 6th, 2006 for del.icio.us and February 16th, 2007 for Hatena Bookmark datasets). firstDate indicates a date when a page had its first bookmark created, hence, when it was added to the social bookmarking system for the first time. It can be seen from Figure 4a that more than a half of the pages were listed among popular URLs in the first three months after being added into del.icio.us. The other half of the pages were bookmarked in the system for the first time more than three months ago. Hatena Bookmark dataset contains even more fresh pages. However, Hatena Bookmark is about one year younger than del.icio.us system. Nevertheless, these figures imply that social bookmarking users often prefer fresh pages. Additionally, almost all pages with PageRank values equal to *0* were posted very recently as it can be seen in Figures 5a and 5b. This last observation suggests that the pages with zero PageRank values are fresh and high-quality pages, which did not have enough time to acquire many inbound links. However, to be completely sure, one would have to know the actual origin dates of these pages[11].

These results highlight one of the useful aspects of SBRank comparing to link-based page ranking metrics. The standard link-based page ranking approach is not effective in terms of fresh information retrieval. This is because pages require relatively long time in order to acquire large number of in-bound links. Consequently, PageRank values of pages are highly correlated with their age. Young pages have difficulties in reaching top search results in traditional search engines even if their

[11] Internet Archive (http://www.archive.org) could possibly provide more constraints on the actual age of the pages. However, we have found that it contains past snapshots of only about 41% of pages from both datasets.

quality is quite high. Figures 6a and 6b show that there are quite low negative values of the correlation coefficients between PageRank and firstDate in our datasets (*r=-0.85* for del.icio.us and *r=-0.51* for Hatena Bookmark datasets). The longer the page existed in the social bookmarking systems, the higher is the probability that its PageRank value is high. We did similar experiment for SBRank vs. firstDate (see Figures 7a and 7b). The correlation coefficient, in this case, had the following values: *r=-0.49* for del.icio.us and *r=-0.08* for Hatena Bookmark datasets.

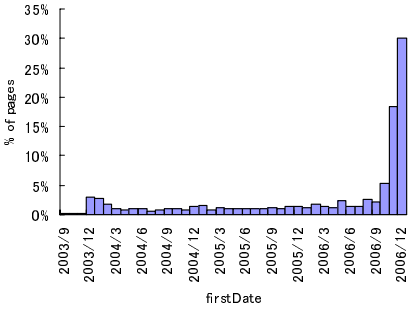

Fig. 4a. Histogram of firstDate of pages (del.icio.us)

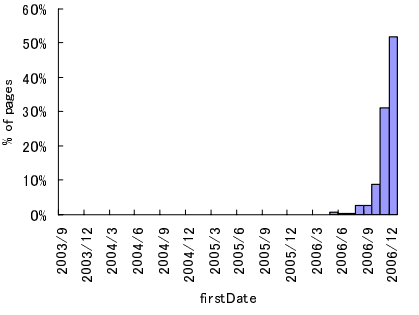

Fig. 5a. Histogram of firstDate of pages that have PageRank value equal to *0* (del.icio.us)

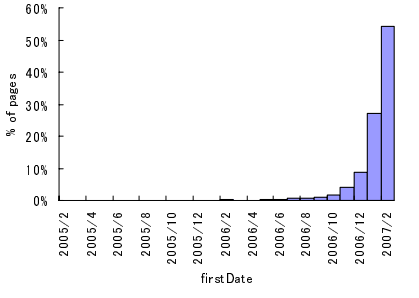

Fig. 4b. Histogram of firstDate of pages (Hatena Bookmark)

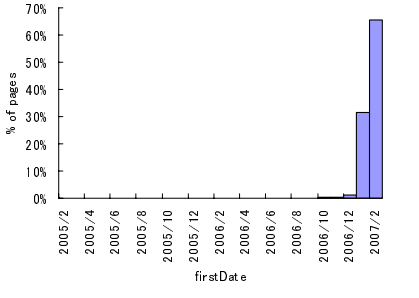

Fig. 5b. Histogram of firstDate of pages that have PageRank value equal to *0* (Hatena Bookmark)

To sum up, the results suggest that SBRank has better dynamics than the traditional link-based page ranking metric. This is because social bookmarks allow for a more rapid, and unbiased, popularity estimation of pages. Complementing PageRank using SBRank has thus potential to bring benefits from the viewpoint of the temporal characteristics of both metrics.

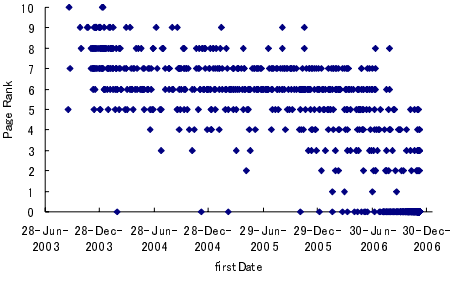

Fig. 6a. Scatter plot of firstDate and PageRank (del.icio.us)

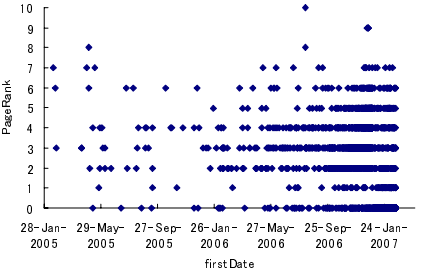

Fig. 6b. Scatter plot of firstDate and PageRank (Hatena Bookmark)

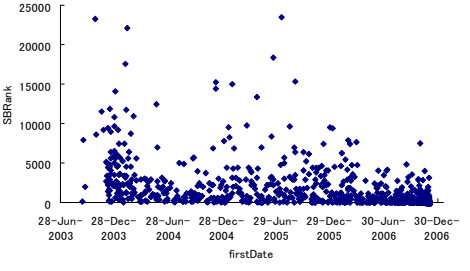

Fig.7a. Scatter plot of firstDate and SBRank (del.icio.us)

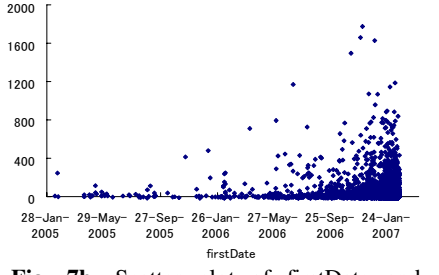

Fig. 7b. Scatter plot of firstDate and SBRank (Hatena Bookmark)

3.5 Hybrid Web Search Proposal

In this section, we demonstrate a simple method for enhancing Web search. Then we show the results of preliminary experiments that were conducted using this method. Such improvement could be simply done by re-ranking top *N* results returned from conventional search engines using the information about the number of their social bookmarks. First, in order to examine whether such an approach would be feasible, we analyzed how many pages in top search results contain at least one social bookmark. Table 1 shows results obtained using Google search engine and del.icio.us social bookmarking system for several sample queries. We can see that, on average, about *79%* of pages returned from Google search engine contain any social bookmarks in del.icio.us and about *23%* in Hatena Bookmark.

To implement a joint rank estimation measure we propose a linear combination of both ranking metrics.

$$CombinedRank_j = \alpha * \frac{SBRank_j}{\max_{\forall j:1 \le j \le N}(SBRank_j)} + (1-\alpha) * \frac{PageRank_j}{\max_{\forall j:1 \le j \le N}(PageRank_j)} \quad (1)$$

$SBRank_j$ is the number of bookmarks of a page j in del.icio.us, while $PageRank_j$ is a PageRank value of the page acquired using Google Toolbar. We normalize both

$SBRank_j$ and $PageRank_j$ values by dividing them by the maximum values found for all N pages. α is a mixing parameter with the value ranging from *0* to *1*.

In the experiment we have used the following queries: "social network", "iphone", "nintendo wii" and "gardening". For each query, we collected $N=50$ top search results from Google search engine. By accessing the returned results, we evaluated the relevance, quality and freshness of each page. Before the manual analysis, pages were randomly ordered to eliminate the potential bias coming from search engine ranking. Quality measures were decided based on several characteristics of pages such as professional outlook, informativeness, text size, number of unique colors and similar features. These characteristics, among others, are usually common for high quality pages [14]. Freshness was determined by analyzing temporal expressions occurring in page content and a general impression of the page's age in case no temporal expressions could be found. Next, we calculated the average value of these three evaluation criteria for each returned page. The resulting values were then used for measuring precision and recall of the results produced by our method.

By applying Equation 1 we could plot precision-recall graphs for each query using different values of parameter α (see Figures 8 to 11). Precision and recall were computed analyzing top k ($k=\{10,20,30,40,50\}$) results within *50* pages returned by the search engine.

From Figures 8-11 it can be seen that PageRank measure used alone ($\alpha=0$) produced better results only for the query "social network" for $k=\{10, 20\}$. On the other hand, SBRank measure used alone ($\alpha=1$) produced the highest quality results for the remaining values of k for query "social network" and for $k=\{20,30,40\}$ for query "iphone". In case of other queries, the hybrid approach was better or at least equally good as PageRank or SBRank measures used alone.

After averaging the precision-recall graphs for all the queries we noticed that the combined approach tends to produce better results for $k=\{10\}$ (Figure 12). On the other hand, there is no improvement of the quality of search results for $k=\{20,30,40\}$ when comparing to PageRank or SBRank used alone.

Choosing the value of the mixing parameter α is a difficult task. As a possible solution, we suggest relating it to one of the two factors, the age of pages or the availability of social bookmarks for pages. Thus, we propose the following three approaches that could be potentially used, in which α_j is a mixing parameter whose value depends on the characteristics of page j:

$$\alpha_j = \frac{1}{t^{now} - t_j^{add}} \tag{2}$$

$$\alpha_j = \frac{1}{t^{now} - t_j^{cre}} \tag{3}$$

$$\alpha_j = \begin{cases} 0 & if \quad PageRank_j > SBRank_j \\ 1 & if \quad PageRank_j \leq SBRank_j \end{cases} \tag{4}$$

Here, t^{now} is the time of query issuing, $t^{add}{}_j$ is the date of the addition of the page j into a social bookmarking system; $t^{cre}{}_j$ is the creation date of the page j. However, detecting creation dates of pages is rather difficult. As a possible solution, creation

dates could be approximated by choosing the minimum value between t^{add}_j and the earliest timestamp of past snapshots of the page j found in any web archive such as the Internet Archive.

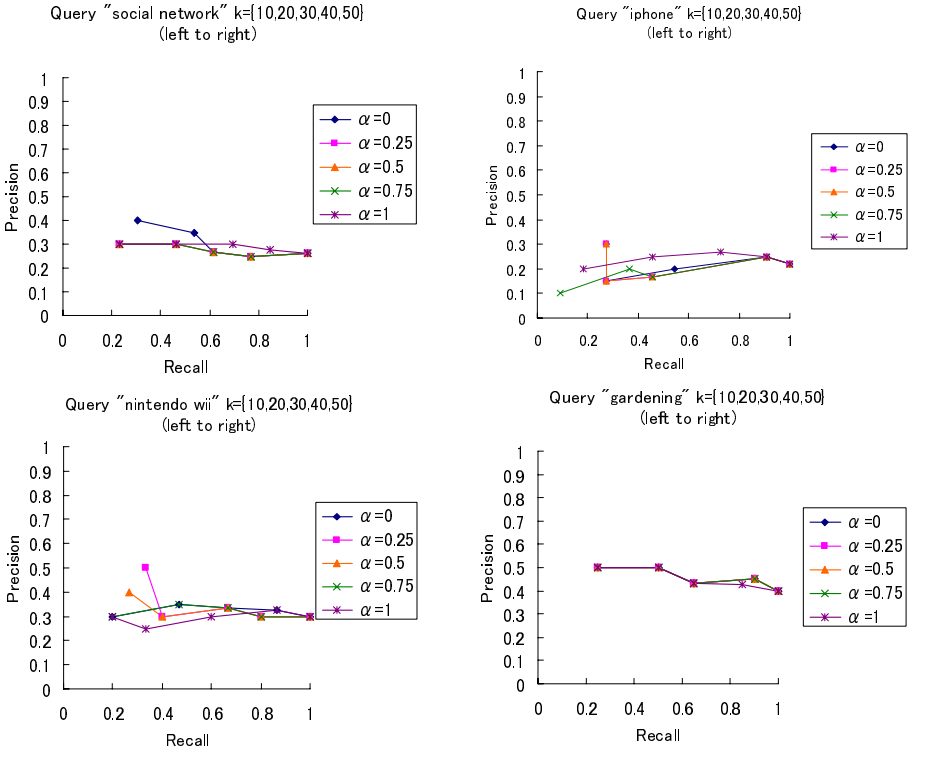

Fig. 8-11. Precision-recall curves for each query: "social network", "iphone", "nintendo wii" and "gardening"

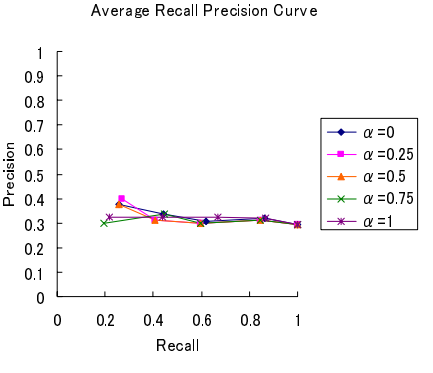

Fig. 12. Average precision-recall curves

Table 1. Number of pages having any social bookmarks for the top 100 results returned from Google search engine for sample queries

query	hatena	del.icio.us	both
graphic design	12	69	69
java	42	92	92
apple	19	83	84
gardening	4	79	79
kyoto	11	61	62
iphone	17	73	74
ipod	37	79	79
steve jobs	26	74	74
ajax	58	86	86
digital library	13	91	91
social network	22	84	84
nintendo wii	14	71	73
Average	***22.92***	***78.5***	***73***

By using Equations 2 and 3, a page would be ranked more by its SBRank measure the younger is its date of insertion into social bookmarking systems or the younger is its actual origin age. This approach would favor the social bookmark-based ranking method in the case of relatively young pages. On the other hand, the approach based on Equation 4 would select the ranking metric that provides a higher value. Thus, pages with few social bookmarks would be ranked more by their PageRank values.

4 Discussion

Web search algorithm that would exploit consumer generated input, which constitutes so-called Web 2.0, is certainly an attractive idea. Several possible directions can be followed to achieve such a "socially-aware" search. For example, swicky[12] is an application that enables building trustful, vertical search engines by communities of users. Our approach is different as we focus on employing social bookmarks made by Web users since they have many advantages over links. Intuitively, there are two main reasons why users create social bookmarks, making pointers to pages for their future reuse or sharing information with other users. It means that either the users expect to revisit bookmarked pages in future or they want to make them known to others. Both objectives allow us to consider social bookmarks as positive votes for pages. Additionally, if we roughly divide Web users into content creators and content consumers, then PageRank can be interpreted as a result of *author-to-author evaluation* of Web resources. On the other hand, SBRank can be considered as a

[12] http://swicki.eurekster.com/

result of *reader-to-author evaluation*. Thus, users who are not capable of creating and managing Web documents could also cast votes for pages leading to a more democratic search process. Another advantage of SBRank over PageRank is that it seems to have better temporal characteristics. SBRank is more dynamic than PageRank, and it often takes short time for pages to reach their popularity peaks in social bookmarking systems [6].

Below, we summarize the observations that we made through our analysis:

- More than half of popular pages in the datasets have lowest PageRank values
  - This implies that many pages which have low PageRank values can be incorporated into top search results through a hybrid Web search
  - It also suggests that people likely discover bookmarked pages from other sources rather than from search engines since many pages in our datasets are relatively difficult to be found by traditional search engines
- Few pages have high SBRank while many pages have rather low SBRank
- There is a weak positive correlation between SBRank and PageRank
  - This result suggests the possibility that SBRank can complement PageRank to enhance Web search
- About half of pages listed as popular in the social bookmarking systems have been introduced in recent three months
  - This indicates high dynamics of SBRank measure and in general of social bookmarking systems as they enable pages to become rapidly popular
  - It also suggests that there are many fresh pages in social bookmarking systems
- There is a high negative correlation between firstDate and PageRank values
  - This result is consistent with the previous observations demonstrating the strong positive bias of PageRank metric towards old pages

In our analysis, we have not considered page relevance that could be estimated by using tags assigned to pages. For example, in the context of link structure analysis, Haveliwala [8] introduced topic-sensitive PageRank. It measures page importance in relation to selected topics, thereby improving page ranking. Similar approach could be adapted to social bookmarks. In this paper, however, we focus on popularity estimation of pages rather than on their relevance.

SBRank is based on user bookmarking activities, however, the importance of each bookmark may be different. A possible extension of our approach would be, thus, to incorporate weighting scheme into SBRank calculation that would depend on the characteristics of users bookmarking pages. This could improve the effectiveness of the page ranking and could help combat potential spamming. Spamming is a threat for every Web search algorithm. Although, until now, no significant spamming attacks have been observed in social bookmarking systems, we think that necessary measures must be taken to prevent deliberate manipulations of social bookmarks in the future. Several measures could be undertaken here as possible lines of defense. For example, user popularity and the history of her or his interactions with the system could be analyzed or users could report suspected inputs themselves.

Lastly, scalability is another problem related to social bookmarking-based Web search. We believe that more data will soon become available along with the

increasing popularity of social bookmarking. Also we hope that efficient meta-search approaches will appear in the near future. In the current situation, we think that the combination of link-based ranking metric and social bookmark-based one is an optimal strategy.

5 Conclusions

Social bookmarks have potential to complement and improve the traditional search in the Web as bookmarked pages are manually checked by multiple Web users, who express their preferences towards pages. Besides improving the quality estimation of pages, social bookmarks can enhance freshness of search results, which is the quality that many search engines currently lack.

In this paper, we investigated the possibility of merging the ranking methods based on the link analysis with the one based on social bookmarks. We have done quantitative studies aiming at comparing both popularity measures and their temporal characteristics. In result of the comparative analysis, we were able to make several observations which allow us to conclude that a hybrid Web search is feasible and useful. We believe that such an analysis is important for the creation of novel search applications considering the weakness of link-based ranking algorithms and the increasing popularity of social collaboration systems in the Web.

In future, we would like to continue the experiments in order to test the proposed approaches. We plan also to work on designing meta-search approaches for improving the search scalability as well as on spam-resistant ranking algorithms.

Acknowledgements

This research was supported by the MEXT Grant-in-Aid for Scientific Research in Priority Areas entitled: Content Fusion and Seamless Search for Information Explosion (#18049041, Representative Katsumi Tanaka), and by the Informatics Research Center for Development of Knowledge Society Infrastructure (COE program by MEXT) as well as by the MEXT Grant-in-Aid for Young Scientists B entitled: Information Retrieval and Mining in Web Archives (Grant#: 18700111), and by "Design and Development of Advanced IT Research Platform for Information" (Project Leader: Jun Adachi, Y00-01, Grant#: 18049073).

References

1. Amitay, E., Carmel, D., Herscovici, M., Lempel, R., Soffer, A.: Trend Detection Through Temporal Link Analysis. Journal of The American Society for Information Science and Technology 55, 1–12 (2004)
2. Baeza-Yates, R., Castillo, C., Saint-Jean, F.: Web Dynamics, Structure and Page Quality. In: Levene, M., Poulovassilis, A. (eds.) Web Dynamics, pp. 93–109. Springer, Heidelberg (2004)
3. Bry, F., Wagner, H.: Collaborative Categorization on the Web: Approach, Prototype, and Experience Report. Forschungsbericht/research report (2003)

4. Cho, J., Roy, S., Adams, R.: Page Quality. Search of an Unbiased Web Ranking. In: Proceedings of SIGMOD Conference, pp. 551–562 (2005)
5. Dwork, C., Kumar, R., Naor, N., Sivakumar, D.: Rank Aggregation Methods for the Web. In: 10th World Wide Web Conference, pp. 613–622 (2001)
6. Golder, S.A., Huberman, B.A.: Usage patterns of collaborative tagging systems, Journal of Information Science, 198–208 (2006)
7. Gomes, D., Silva, M.J.: Modeling information persistence on the Web. In: Proceedings of the 6th International Conference on Web Engineering, Palo Alto, CA, USA, pp. 193–200 (2006)
8. Haveliwala, T.H.: Topic-Sensitive PageRank: A Context-Sensitive Ranking Algorithm for Web Search. IEEE Transactions on Knowledge and Data Engineering, 784–796 (2003)
9. Keller, R.M., Wolfe, R.R., Chen, J.R., Labinowitz, J.L., Mathe, N.: A Bookmarking Service for Organizing and Sharing URLs. In: Proceedings of the 6th International World Wide Web Conference, Santa Clara, CA, pp. 1103–1114 (1997)
10. Kleinberg, J.M.: Authoritative Sources in a Hyperlinked Environment. Journal of the ACM, 604–632 (1999)
11. Lawrence, S., Giles, C.L.: Inquirus, the NECI meta search engine. In: Proceedings of the 7th International World Wide Web Conference, Brisbane, Australia, pp. 95–105 (1998)
12. Liu, B.: Web Data Mining: Exploring Hyperlinks, Contents and Usage Data. Springer, Heidelberg (2007)
13. Lu, Y., Meng, W., Shu, L., Yu, C., Liu, K.: Evaluation of Result Merging Strategies for Metasearch Engines. In: Proceedings of Web Information Systems Engineering conference, pp. 53–66 (2005)
14. Mandl, T.: Implementation and evaluation of a quality-based search engine. Hypertext2006, pp. 73–84 (2006)
15. Marlow, C., Naaman, M., Boyd, D., Davis, M.: HT06, Tagging Paper, Taxonomy, Flickr, Academic Article, To Read. Hypertext2006, pp. 31–40 (2006)
16. Mathes, A.: Folksonomies - Cooperative Classification and Communication Through Shared Metadata. Computer Mediated Communication, LIS590CMC (2004)
17. Ntoulas, A., Cho, J., and Olston, C.: What's new on the Web? The evolution of the Web from a search engine perspective. In: Proceedings of the 13th International World Wide Web Conference, New York, USA, pp. 1–12 (2004)
18. Page, L., Brin, S., Motwani, R., Winograd, T.: The pagerank citation ranking: Bringing order to the Web. Technical report, Stanford Digital Library Technologies Project (1998)
19. Strutz, D.N.: Communal Categorization: The Folksonomy. INFO622: Content Representation (December 2004)
20. Wu, X., Zhang, L., Yu, Y.: Exploring Social Annotations for the Semantic Web. In: World Wide Web Conference, pp. 417–426 (2006)
21. Wu, H., Zubair, M., Maly, K.: Harvesting Social Knowledge from Folksonomies. Hypertext2006, pp. 111–114 (August 2006)
22. Zhang, L., Wu, X., Yu, Y.: Emergent Semantics from Folksonomies: A Quantitative Study. In: Spaccapietra, S., Aberer, K., Cudré-Mauroux, P. (eds.) Journal on Data Semantics VI. LNCS, vol. 4090, pp. 168–186. Springer, Heidelberg (2006)
23. Yu, P.S., Li, X., Liu, B.: On the temporal dimension of search. In: Proceedings of the 13th International World Wide Web Conference, New York, USA, pp. 448–449 (2004)