Can Social Bookmarking Enhance Search in the Web?

Yusuke Yanbe Adam Jatowt Satoshi Nakamura Katsumi Tanaka

Department of Social Informatics, Kyoto University
Yoshida-Honmachi, Sakyo-ku
606-8501 Kyoto, Japan
Phone: +81-75-753-5969
{yanbe, adam, nakamura, tanaka}@dl.kuis.kyoto-u.ac.jp

ABSTRACT
Social bookmarking is an emerging type of a Web service that helps users share, classify, and discover interesting resources. In this paper, we explore the concept of an enhanced search, in which data from social bookmarking systems is exploited for enhancing search in the Web. We propose combining the widely used link-based ranking metric with the one derived using social bookmarking data. First, this increases the precision of a standard link-based search by incorporating popularity estimates from aggregated data of bookmarking users. Second, it provides an opportunity for extending the search capabilities of existing search engines. Individual contributions of bookmarking users as well as the general statistics of their activities are used here for a new kind of a complex search where contextual, temporal or sentiment-related information is used. We investigate the usefulness of social bookmarking systems for the purpose of enhancing Web search through a series of experiments done on datasets obtained from social bookmarking systems. Next, we show the prototype system that implements the proposed approach and we present some preliminary results.

Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *Query formulation, Retrieval models, Search Process*

General Terms
Algorithms, Experimentation, Theory

Keywords
social bookmarking, social search, PageRank, metadata

1. INTRODUCTION
Information retrieval (IR) has the objective of obtaining relevant documents from document collections given queries provided by users. Traditionally, vector space model based on popular TF*IDF measure [19] was used for finding relevant documents. This approach works well in finite and controlled environments like document collections. However, in huge and uncontrolled environments like Web, a simple content-based retrieval method such as the vector space model is impractical. In the Web, large quantities of resources constantly compete for the attention of users. Many of them are indeed relevant to user queries (which are usually very short and often ambiguous). However, the sheer size of the Web does not allow for presenting the entire set of related documents to users. In such an environment the quality of documents starts to play an important role, yet measuring this quality basing solely on page content is difficult and may be also subjective.

The link structure of the Web provides better means for estimating page qualities. PageRank [18] is the most famous method that uses link structure analysis. The idea behind PageRank algorithm is to exploit the macro-scale link structure between pages in order to capture the popularity of documents and indirectly their qualities. According to this approach, the popularity of a page is determined on the basis of the size of a hypothetical user stream coming to the page. However, link-based algorithms have currently many disadvantages [14], for example they are vulnerable to spamming, it is often difficult to create links for average users, links may have variety of meanings and purposes, etc. Therefore, despite the previous success of link-based search algorithms, their current limitations cause that new, better approaches need to be sought.

With the advent of Web 2.0, social bookmarking systems seem to have a potential for improving search capabilities of current search engines. In these systems, the popularity of a Web page is calculated as the total number of times it has been bookmarked, hence, by the number of users voting for the page. We call this measure SBRank. There are many differences between PageRank and SBRank that result from their characteristics. SBRank captures the popularity of resources among content consumers (page readers), while PageRank is in general a result of author-to-author evaluation of Web resources. This means that users who are not capable of creating and managing Web pages could also give "votes" to pages by creating social bookmarks. This situation is probably one of the causes of different temporal characteristics of both metrics. Generally, SBRank is more dynamic than PageRank, and it often takes short time for pages to reach their

JCDL'07, June 18–23, 2007, Vancouver, British Columbia, Canada.

popularity peaks in social bookmarking systems [9]. Besides this, an important advantage of social bookmarking systems is metadata that is associated with resources. Tag patterns emerge for documents tagged by multiple users allowing for describing their contents and characteristics. This contextual information is created by content consumers (reader-to-author evaluation); hence it should be more trustful than the metadata provided by content producers.

However, despite many advantages of social bookmarking services, relying on them alone is currently still not possible in a general Web search. This is due to the insufficient amount of bookmarked pages available for any arbitrary query to generate satisfactory results. We are recently observing a rapid increase in the number of bookmarked pages, yet the combination of link structure-based and social bookmarking-based page ranking measures seems to be currently an optimal strategy.

In this paper, first we attempt to make a comparative analysis between PageRank and SBRank ranking metrics. The objective of this investigation is to analyze the potential of a hybrid Web search method that would use both measures for improving the efficiency of quality estimates used by current search engines. For this purpose, we have conducted several analytical studies based on data retrieved from social bookmarking systems, which confirmed the feasibility and effectiveness of the proposed approach.

As a second contribution of this paper, we propose using data from social bookmarking systems for enabling a new kind of search. Since pages in social bookmarking systems are freely annotated with tags describing their contents, then a search based on the contributions of multiple users can be performed. This type of a search is similar to the one based on metadata in digital libraries or other collections of annotated resources. The usual content-based search approach cannot capture page content and characteristics well, unless the information about them is explicitly provided by authors inside document content. We propose here utilizing consumer-provided metadata for extending capabilities of current search engines.

Besides the improved relevance, tags allow for a more complex quality estimation of pages. This can be achieved by using sentiment tags, user comments and, in general, global statistics derived from user behavior in relation to pages. For example, it is possible to search for pages that feature certain characteristics like being useful or funny. This type of sentiment analysis is not feasible using only page content or standard popularity rankings. Additionally, social bookmarking systems allow for temporal search since bookmarks usually have timestamps provided. For example, it is possible to distinguish pages that are fresh from obsolete ones or to detect pages with certain popularity patterns.

We describe the prototype system that we have implemented for exploring the usefulness of our approach. It combines both PageRank and SBRank metrics as well as uses user-created annotations and general statistics of user behavior towards pages. In result, our application allows users to search for Web pages by their content, associated metadata, temporal aspects in social bookmarking systems, user sentiment and other features. We show the results of preliminary experiments to demonstrate the attractiveness of the proposed search type.

The remainder of this paper is organized as follows. Section 2 provides the necessary background and Section 3 discusses the related research. Section 4 demonstrates the results of the analytical studies that we made. In the next section we introduce our approach for a new, enhanced search type. In Section 6 we demonstrate a prototype system and show preliminary experiments. Lastly, Section 7 concludes the paper and provides a brief look at our future work.

2. BACKGROUND

2.1 Link-based Page Ranking

Link-based page ranking algorithms and especially PageRank are considered to be the driving force of current search engines. In PageRank method, the popularity of a page is determined basing on a random surfer model, where the probability of the surfer reaching a certain page is calculated as the result of a random selection of links on visited pages. As mentioned above, link-based page ranking approach has certain disadvantages. One is related to the relative difficulty of creating links. Making hyperlinks requires some effort for users as they have to create pages. Page authors are also limited to the topical scopes of pages and thus cannot freely insert any arbitrary link they want; as such an insertion would have to fit into the document content and its topics. Although, recently we observe the explosion of blogs, which make the link creation relatively easy; yet many search engines started the policy of neglecting or negatively biasing links from blogs due to spamming problems. In general, links are created by Web authors who produce and maintain content in the Web. However, there is an overwhelming group of Web users who are non-authors, and whose voices and opinions are not exploited by the standard link-based ranking metric.

Moreover, along with the development of Web-related technology many applications appeared that can generate hyperlinks automatically. This also raises the question of the actual value of such hyperlinks. Lastly, spam trackback, splog (an automatically generated spam blog), and other spam techniques currently pose threat to the effectiveness of link-based algorithms [8]. In general, the link analysis approach is still useful, but we believe that it needs to be complemented by another reliable metric.

2.2 Social Bookmarking

Manually filtered page collections are usually of high quality and provide trustful information sources. In the early years of the Web, directory services were utilized in order to arrange the Web. These were managed by professional editors, who manually selected useful resources. However, the rapid growth of the Web soon made this approach impractical.

In 2003 Joshua Schachter launched del.icio.us[1] - the first social bookmarking service. Inspired by del.icio.us, many kinds of social bookmarking systems have been established recently. The simplicity they offer for creating bookmarks and adding annotations was one of the reasons for their high popularity. The advantage of social bookmarking systems over Web directories comes mostly from the fact that bookmarking and tagging is useful for individual users who want to externally store access points to their selected resources. This means that users have an immediate profit from their actions, by which they also indirectly

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

help to manually arrange the Web in a bottom-up fashion. The advantage of social bookmarking systems is that 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 many bookmarks. Users can also subscribe to "Inbox" a bookmark activity reporting service. From this feedback, useful pages attracting much attention can become rapidly known to many users.

An important characteristic of tagging in social bookmarking systems is that there is no controlled vocabulary used. Rather tags are added freely and agreements emerge among users, who learn the ways others tag and describe resources [9]. Thus, tags are different from the professional and rigid classification done, for example, by librarians. Recently, folksonomy appeared as the name of user created uncontrolled tag collections. Although, ambiguity and synonymy are the main problems related to folksonomy, yet, the freedom of tag choice and the lack of taxonomy-related knowledge necessary to be acquired make tagging a popular activity.

There are also social bookmarking services for other digital resources than Web pages. CiteULike[2] is an example of a social bookmarking system that allows users to share information about scientific papers. Traditionally, the impact of research publications is measured by the number of their citations. By using the numbers of social bookmarks and the associated with them tags it is possible to obtain another metric for estimating the popularity and quality of publications.

3. RELATED WORK

3.1 Web Search

Previous attempts at arranging the Web and making it accessible to users were based on cataloging and using standard IR techniques. Since these approaches became impractical due to the rapid growth of the Web, several attempts that exploit link structure of the Web were proposed.

The above introduced PageRank is the most famous link-based page ranking algorithm currently used in the Web. HITS [13] is another well-known link-based algorithm. It computes two types of page quality estimates, hub score and authority score. Authority score is a measure based on the number of links coming from pages that have high hub scores. Hub score, on the other hand, measures how many links are from pages with high authority scores. The calculation of both PageRank and HITS is done recursively until the algorithms converge to stable results. Topic-Sensitive PageRank proposed by Haveliwala et al. [12] is an extension of PageRank. It incorporates page relevance into the standard PageRank measure by computing sets of topic-biased scores for pages. The authors demonstrate higher effectiveness of the topic-specific approach over the standard PageRank algorithm.

Much effort concentrated previously on utilizing global statistics of users and their behavior for selecting high quality and relevant resources. For example, search engine log analysis [3,6] enabled to associate pages with queries by investigating the frequencies of users accessing pages from returned results. This could be viewed as a metadata constructing approach, in which large numbers of users indirectly help by their actions to describe page content. Alexa[3] is an interesting example of a search engine that provides search capability based on collecting large scale statistics about users visiting pages in real time. This is possible thanks to collecting data from many Web browsers. In result, an accurate and timely estimation of page popularities is possible. On the other hand, semantics emerging from user browsing paths in multimedia collections were exploited in [11,17]. Our work is similar to these efforts in the sense that we use information about general user behavior towards given resources. This is not only agglomerated information about tags but also statistics on global changes of user preferences and actions in time.

Temporal link analysis [1,2,5] focuses on detecting link evolution, link change patterns or on utilizing time-related information of links for improving page ranking. The method for finding authority pages in selected time frames as well as detecting trends in the Web was introduced in [1]. Cho et al. [5] proposed a model for estimating quality of pages by analyzing the dynamics of in-bound links as an alternative to static popularity-based ranking. Pages that have growing trends of popularity (large increases of in-bound link numbers), while being still relatively unpopular in the Web, were considered to have highest qualities. This approach may be, however, difficult to use in practice due to the lack of available data about the link structure from the past. It should be noted that social bookmarking systems provide timestamps attached to links. Thus, in contrast to link-based ranking, incorporating temporal aspects into the search is generally feasible here.

In other related studies, Baeza-Yates et al. [2] suggested modifying PageRank computation by incorporating last-modification dates of pages. [24] proposed Timed PageRank algorithm based on exponentially decaying PageRank scores of linking pages. These approaches were motivated by the observation that PageRank is biased against new pages as it takes some time until pages become noticed and trusted by Web authors [2,10].

3.2 Social Bookmarking Research

Although, already some analyses have been done on social bookmarking [4,9,15,16,20,21,22,23], however, it still has not been sufficiently studied until now. Previous studies focused mostly on the issues related to folksonomy [20,21,22,23]. For example, Zhang et al. [23] 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. Golder and Huberman [9] investigated the nature of tagging and bookmarking using data obtained from del.icio.us. They discovered interesting regularities in user activities, tag frequencies, and bursts in popularity of tags in social bookmarks. The authors also analyzed tagging dynamics as well as classified tags into seven categories depending on the functions they perform for bookmarks. In another work, Marlow et al. [15] proposed a general taxonomy of tagging systems and user incentives in social bookmarking. None of the previous studies, however, focused on comparative analysis of link structure and social bookmarking metrics neither on the possibility of exploiting social bookmarks for enhancing Web search.

[2] http://www.citeulike.org

[3] http://www.alexa.com

Since social bookmarks have rather complex data structure (user, resource, tag), building an efficient search model is a challenging task. In [22] HITS algorithm was adapted for identifying high quality resources and users that provide such resources in the Web. In another work, Damianos et al. [7] proposed using social bookmarking for information sharing and management in corporate Intranets. Finally, Wu et al. [21] described techniques for exploiting social bookmarking for the purpose of fostering the development of Semantic Web. The authors used probabilistic generative model to capture emerging semantics of resources.

In contrast to the previous approaches, we focus on merging link-based ranking metrics with the metric that leverages results of collaborative tagging. Additionally, we propose exploiting other characteristics of social bookmarking systems such as general user behavior, sentiment of users towards bookmarked resources, the stimulus levels that resources cause, etc. These extensions enable to achieve a complex search mechanism that offers interesting search capabilities.

4. ANALYTICAL STUDY

In this section, we present and discuss results of analytical studies done in order to compare SBRank and PageRank metrics, as well as to examine the usefulness of social bookmarking for enhancing search in the Web.

4.1 Datasets

We have chosen del.icio.us as a source of our dataset since it is currently the most popular social bookmarking service for Web pages. Related previous studies have also used data from this service [9,23]. We have utilized *popular tags*[4], which is a set of the most popular and recently used tags that are continuously published by del.icio.us. In total 140 tags were obtained on December 6th, 2006. In the next step, 2,673 popular URLs were retrieved using these tags. After removing duplicate URLs, which were listed under several tags, we obtained 1,290 unique pages. Each page had the following attributes: Tags, URL, FirstDate and SBRank. FirstDate indicates the time point when the page was introduced to the system for the first time by being bookmarked by one of users.

Next, we retrieved PageRank values of the pages using Google Toolbar[5]. Google Toolbar[6] is a browser toolbar that allows viewing PageRank values of accessed pages. PageRank values are approximated on the scale from 0 to 10 (0 means the lowest PageRank value).

In Section 4.4, we use another dataset that was collected from Hatena Bookmark[7] - the most popular social bookmarking service in Japan. The data was retrieved from the archive of the listings of popular pages called "hot entry"[8] for the period of 100 days from September 14, 2006 to December 23, 2006. We obtained 50 URLs for each day. After removing duplicates, 3,663 unique URLs were left, for which bookmark data was collected. Since the dataset contained most of tags written in Japanese, we have translated them into English.

4.2 Distributions of PageRank and SBRank

Figure 1 shows the percentage distribution of PageRank values in del.icio.us dataset. It is interesting to observe that there is a high number of pages with low PageRank values (56.1% of URLs have PageRank value equal to 0). One conclusion that comes from this result is that finding these pages using conventional search engines is relatively difficult. Pages with low PageRank values have difficulty reaching top places in search results even if their relevance is high.

Next, we plotted the distribution of SBRank values for the pages from the first dataset (Figure 2). We could observe that they were bookmarked on average by 50 users. Hence, the pages were considered as being valuable for a relatively large number of users. Figure 2 shows that a few pages (top 10%) are bookmarked by many users, while the rest are bookmarked by a relatively low number of users. The median number of bookmarks is 144, while the average is equal to 1115.

Considering the high popularity of pages among del.icio.us users and their low PageRank values, it seems that social bookmarking users are good at gathering high quality pages that were not discovered yet by the wider audience. However, it may also imply that pages were found by users from other sources than conventional Web search engines (possibly by interacting with social bookmarking systems).

Figure 3 displays a scatter plot of both measures. It indicates a weak positive correlation coefficient (r=0.53) between both SBRank and PageRank values. Note that 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 complementing it with SBRank. On the other hand, if correlation coefficient between both metrics had a very low value, it would mean that one of them is wrong. Since the result is within the acceptable range, we can consider combination of both quality estimates to be possible.

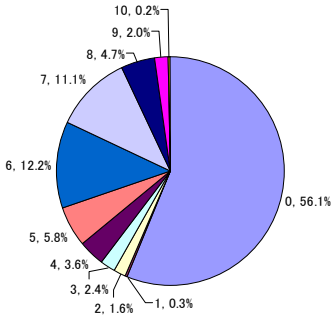

Fig. 1. Distribution of PageRank values

[4] http://del.icio.us/tag

[5] As Google API does not provide any automatic method for acquiring PageRank scores they were manually collected using Google Toolbar

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

[7] http://b.hatena.ne.jp

[8] http://b.hatena.ne.jp/hotentry

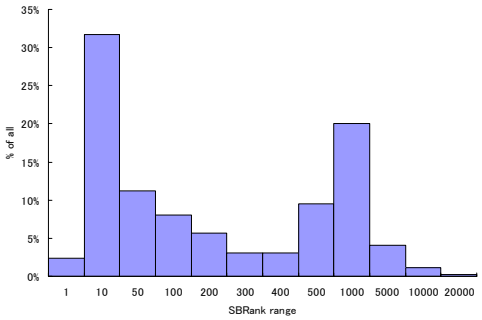

Fig. 2. Histogram of SBRank

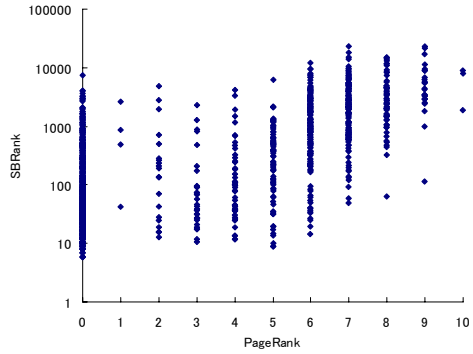

Fig. 3. Scatter plot of PageRank and SBRank (logarithmic scale)

4.3 Temporal Analysis

In this section, we focus on temporal aspects of social bookmarking systems. Figure 4 shows the distribution of FirstDate (dates of addition) for the pages. We can see that about half of the pages were listed among popular URLs in the first three months since their addition into del.icio.us. The other half were bookmarked for the first time more than three months before the data collection date. This implies high dynamics of bookmarking systems in general. Although this result depends on del.icio.us's definition of popular tags, it is still interesting as an indicator of a typical behavior of users in social bookmarking systems.

Figure 5 displays the same distribution for pages with PageRank values equal to 0. As it was shown above, such pages represent over half of the documents in the dataset. On average, they were introduced to the social bookmarking system more recently than the rest of the pages. This can be found after comparing Figures 4 and 5. This result confirms the previous observation made by Golder and Huberman [10] who reported that 67% of pages reached their peak popularity levels in the first 10 days after being added to del.icio.us.

On the other hand, 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 [2,5,10]. Consequently, young pages cannot be found through traditional search engines even if their quality and relevance are quite high. This negative bias towards young pages was observed by Baeza-Yates et al. [2] in the study on Chilean Web sites who noted that the PageRank values are highly correlated with their age. To some extent Figure 6 supports this observation showing that FirstDate and PageRank values reveal quite high negative correlation coefficient (r=-0.85).

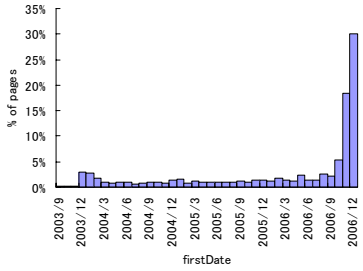

Fig. 4. Histogram of FirstDate of page

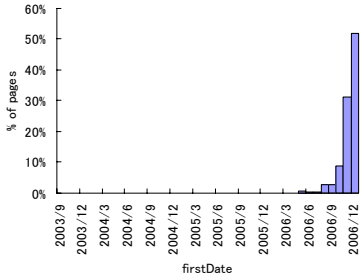

Fig. 5. Histogram of FirstDate of pages that have PageRank value equal to 0

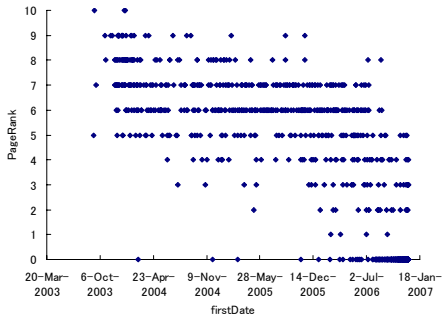

Fig. 6. Scatter plot of FirstDate and PageRank

4.4 Sentiment Analysis

One of interesting characteristics of social bookmarking is that often tags contain sentiments expressed by users towards bookmarked resources. This could allow for a sentiment-aware search that would exploit user feelings about Web pages. The merit of a sentiment-based quality estimate of Web documents cannot be exaggerated as many times users require resources that reveal certain sentimental characteristics: for example pages that are funny, useful or inspirational.

In order to measure the number and kinds of sentiment tags used by bookmarking users we have used the second dataset, which was obtained from Hatena Bookmark. Tags in this dataset were classified into two groups according to tag taxonomy defined by Golder and Huberman [9]:

a) tags that identify what or whose the resource is about

b) tags that identify qualities or characteristics of resources (scary, funny, stupid etc.)

We call these tag types content and sentiment tags, respectively. We have manually examined top 1,100 tags from our dataset to detect content and sentiment tags. Then, we have translated them into English. Table 1 shows top 10 content and sentiment tags. In some cases, the same tag is listed several times, since in Japanese there are often several words used to express the same meaning. Nevertheless, it can be seen that content tags are on average more common than sentiment tags. We observed that in the top 30 tags the ratio of content tags to sentiment tags is about 10:1. In Figure 7 we show the distribution of tag frequencies. Top 3 sentiment tags are very common, while the other tags are rather less used. After including synonyms we found that the most popular sentiment tags are: useful, amazing and awful.

Figure 8 presents top 54 sentiment tags placed on the negative-positive scale including the information about their frequencies. The tags appearing more than 3000 times are above the dashed line, while those with frequencies less than 100 times are below the horizontal axis. In general, there are more positive sentiment tags than negative ones and positive ones are also more frequently used. Only one negative sentiment tag was used more than 100 times ("it's awful"). This means that social bookmarkers usually do not bookmark resources to which they have negative feelings.

Table. 1. Top 10 content tags (left) and top 10 sentiment tags (right)

Tag Name	N
Web	16,633
google	15,674
troll	14,453
javascript	11,840
youtube	10,858
tips	10,784
css	9,411
design	8,423
2ch (huge BBS)	8,381
society	7,412

Tag Name	N
useful (1)	5,381
it's amazing	5,046
it's awful	4,123
useful (2)	3,041
interesting	638
funny (1)	616
it's useful (3)	544
funny (2)	419
useful (4)	377
I see	365

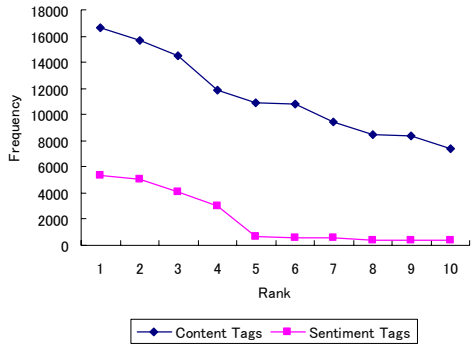

Fig. 7. Frequency distribution of top 20 content and sentiment tags

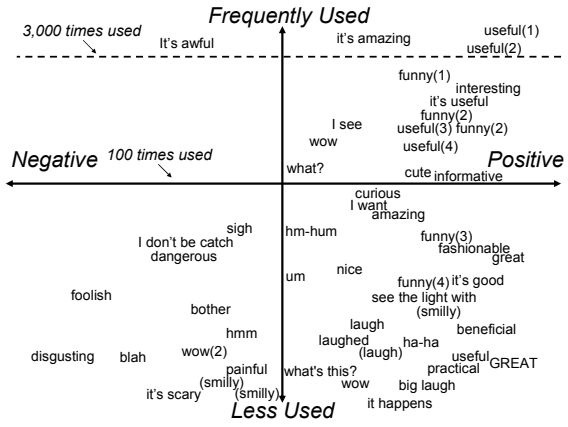

Fig. 8. Top 54 sentiment tags on the sentiment scale

5. ENHANCED SEARCH

In this section we demonstrate our proposal for constructing an efficient search system with extended search capabilities. We discuss basic elements of a complex query that can be constructed as well as the algorithm that handles such queries.

5.1 Complex Query

5.1.1 Enhancing Page Popularity Ranking

Complementing PageRank and SBRank seem to be useful in the light of the experiments that we have previously demonstrated. The high dynamics of social bookmarking services makes it superior over traditional link based page ranking metric from the temporal viewpoint, as it allows for a more rapid, and unbiased, popularity increase of pages. This was showed in Section 4.3. Also, in the experiments in Section 4.2, many pages which have relatively low PageRank values were found among popular pages in the social bookmarking system. Such pages should have their ranks increased so they can be included into top search results of a hybrid Web search. We propose a straightforward way of merging both metrics using a linear combination.

5.1.2 Metadata Search

Tags annotated by users are useful for a so-called search by metadata ("metadata query"). Wu et al. [21] introduced a search model that exploits tags of bookmarked pages. Here, in contrast to that proposal, we focus on using tags to enhance capabilities of existing Web search engines.

It was shown in [9] that tag proportions stabilize after certain time. This means that bookmarking users reach agreement on the kind of tags to be used for describing resources. We assume that this process takes place because users observe tags made by others.

In our search model a user can issue both a traditional query, which is called "content query" as well as "metadata query" at the same time. In such a case, pages that contain content query in their contents will have their ranks computed considering the mapping of metadata query and the tags describing the documents. We construct a tag vector that represents tags associated with the page and their frequencies. The similarity between metadata query and the tag vector is the basis for assigning scores to pages.

5.1.3 Temporal Query Support

Temporal queries can be constructed exploiting timestamps associated with bookmarks. This type of search cannot be easily

realized in traditional link-based approaches, since no data is currently available on the link evolution of the Web.

First, we propose filtering pages according to FirstDate values, that is, according to the freshness levels of pages inside social bookmarking systems. This allows for improving fresh information retrieval. Naturally, pages could have been created some time ago before being noticed and bookmarked. However, we think that the capabilities of social bookmarking systems to produce more fresh results will increase in the future. The growing number of social bookmarking users and the potential for merging data from different systems are the bases of this belief.

As the next ranking criteria, we use the variance measure of the function representing changes in SBRank in time. The reason behind this choice is the objective of capturing simple temporal patterns of page popularity in time. For example, search can be done for pages with stable popularity function or for pages with high peaks reflecting large changes in user preferences in time.

Lastly, we propose the measure for capturing levels of page popularities in certain, specified periods of time. This can be estimated by summing the numbers of bookmarks made to documents during user-selected intervals. Queries of the type: "pages that were popular from t_{beg} to t_{end}" can be thus constructed.

5.1.4 Sentiment Query Support

As mentioned in Section 4.4, about 10% of tags used in social bookmarking systems are sentiment type tags. We can use them for implementing sentiment-based search. For example, a user may request a page about Kyoto that is interesting or useful.

Page sentiment vector is created considering the sentiment tags added by users. The similarity of this vector with user-issued sentiment queries is then taken to compute sentiment-based scores of pages.

5.1.5 Controversial Query Support

Social bookmarkers tend to leave comments in bookmarks when pages have controversial contents. In the context of explorative search, users sometimes want to search for opinions or discussion about certain topics. We can thus add another aspect of search to use numbers of comments on pages added by social bookmarkers.

5.2 Algorithm

At query time the system performs the following operations:

1. Obtain top n pages from search results returned by a search engine $P=\{p_1, p_2, \ldots, p_n\}$ for query q
2. Obtain *SBRank* values for each p_i where $p_i \in P$
3. Obtain every bookmark and its associated data for each p_i that has $SBRank > 0$ (i.e., the page has at least one social bookmark)
4. Count the number of occurrences of users and tags to be used for providing "Related Tags" and "Related Bookmarks" capabilities (described in Section 6.1)

In order to incorporate query types described in Section 5.1, we have applied the ranking formula shown in Figure 9. The original search results returned by the search engine are re-ranked using $Rank(p_i)$ function.

$$Rank(p_i) = (1+B(p_i)) \cdot (1+F(p_i)) \cdot (1+V(p_i)) \cdot (1+C(p_i)) \cdot (1+T(p_i,q)) \cdot (1+T^{sen}(p_i,q)) \cdot (1+S(p_i,t_{beg},t_{end}))$$

where:

$$B(p_i) = \alpha \cdot SBRank(p_i) + (1-\alpha) \cdot SearchRank(p_i)$$

$$F(p_i) = \beta \cdot \frac{FirstDate(p_i) - \min_{1 \le j \le n}(FirstDate(p_j))}{\max_{1 \le j \le n}(LastDate(p_j)) - \min_{1 \le j \le n}(FirstDate(p_j))}$$

$$V(p_i) = \gamma \cdot \frac{Var(p_i, FirstDate(p_i), LastDate(p_i))}{\max_{1 \le j \le n}(Var(p_j, FirstDate(p_j), LastDate(p_j)))}$$

$$C(p_i) = \delta \cdot \frac{N(comment_i)}{N(comment)}$$

$$T(p_i,q) = \alpha \cdot sim(tag_i, tag_q)$$

$$T^{sen}(p_i,q) = \alpha \cdot sim(tag^{sen}{}_i, tag_q)$$

$$S(p_i,t_{beg},t_{end}) = \frac{AddBook(p_i,t_{beg},t_{end})}{AddBook(p_i,FirstDate(p_i),LastDate(p_i))}$$

$SearchRank(p_i)$ is a rank of p_i

$SBRank(p_i)$ is the number of bookmarks of p_i

$FirstDate(p_i)$ is the first date when a bookmark was made to p_i

$LastDate(p_i)$ is the last date when a bookmark was made to p_i

tag_i is a tag vector of p_i

$tag^{sen}{}_i$ is a sentiment tag vector of p_i

tag_q is a tag vector of q

$AddBook(p_i, t_{left}, t_{right})$ is the number of bookmarks-made to p_i in (t_{left}, t_{right})

$sim(tag^{sen}{}_i, tag_q)$ is similarity between tag vector of p_i and tag vector of q

Fig. 9. Ranking Formula

Below we describe the symbols used in the algorithm. $B(p_i)$ represents the popularity estimate of p_i using the combination of $SBRank(p_i)$ and $SearchRank(p_i)$, which is the rank of the page in the results returned from a search engine. $F(p_i)$ is the freshness level of p_i; $V(p_i)$ is a variance measure of the function representing added bookmarks to p_i; $C(p_i)$ reflects the controversy level of page p_i expressed as the number of its comments, $N(comment_i)$. $N(comment)$ is the highest number of comments for all pages in P. $sim(tag_i, tag_q)$ is the similarity between page tag vector and query vector, while $sim(tag^{sen}{}_i, tag_q)$ is the similarity between the page sentiment vector and the query vector. $S(p_i, t_{beg}, t_{end})$ is the proportion of bookmarks of p_i, which have been added in the time period $<t_{beg}, t_{end}>$ to the total number of bookmarks added to this page in the bookmarking system. Lastly, α, β, γ and δ are controlling parameters.

6. EXPERIMENTS

6.1 System Implementation

We have implemented a prototype of the system that provides the enhanced search capabilities described above. The snapshot of the system's interface is shown in Figure 10. The system was implemented using Python and JavsScript programming languages used for rapid and dynamic prototyping, and for enabling interactive functionalities. It was deployed on Intel Xeon™ CPU 3.20GHz server with 8GB physical memory. To improve response time and to reduce the cost of accessing social bookmarking systems, bookmark data was cached on a local hard disk for 1 week. We have used Hatena Bookmark as the data source.

The prototype application has a graphical user interface that allows issuing complex queries in an easy and intuitive way. Slide-bar control interfaces were added for adjusting α, β, γ and δ parameters according to user queries. "Time span" control was implemented as an interactive interface with two sliding bars allowing for specifying time periods. User could also introduce the limits of the desired time span into textual boxes. Radio buttons were used to choose among three most popular sentiment expressions that were identified in Section 4.4: useful, amazing and awful. Only three sentiment queries were chosen here, since it is difficult to integrate all potential sentiment tags into one interface. We have manually made a simple sentiment dictionary, in which related sentiment expressions or synonyms had weights assigned to be mapped into the three basic sentiment categories.

Normally, all controls are at their default levels, that is, at the positions, when they do not influence search results. Thus, user can make the standard content-based search without using any additional query features.

For facilitating the usage, the system allows for issuing queries expressed fully in text without the need for using GUI interface. The query patterns to be used for this purpose are listed below:

- search:x where x is a content query
- SBRank:α
- tag:z where z is a metadata query
- freq:(peaky|cont)
- time:(old|new)
- (from|to):YYYYMMDD
- emo:(useful|amazing|awful)
- res:(buzz|quiet)
- lang:(ja|en|all)

Additionally, the system dynamically generates navigable structures called "related tags" and "related bookmarks" according to issued queries for enabling serendipitous discovery (see Figure 10).

"Related tags" area displays 20 most frequently occurring tags for all pages returned in results. Tags are displayed using varying font sizes according to their frequencies. Users can then explore other tags related to issued queries. After clicking on tags, pages containing tag-related information appear from social bookmarking systems, while after clicking on the "+" signs associated with related tags, new search queries containing those tags are issued. The settings of the other query parameters remain unchanged then.

"Related bookmarks" are tuples of social bookmark users and their tags calculated using the search results. If many pages returned in search results were bookmarked by the same user and they also achieved high ranks inside these results then the weight of the user is increased. For top-scored users, their most frequent tags that are also common to the pages returned in results are displayed together with links to their corresponding pages in social bookmarking systems.

6.2 Evaluation

We show here preliminary evaluation of the proposed search enhancement. Several complex queries were issued to the system. In Table 2 we show page titles of top 3 results for each query together with their original ranks assigned by Google search engine. Left column shows queries that were used indicating those query features that had different values than default ones. Queries were translated into English.

Our system returned pages titled "Internet Archive", "CiteSeer" and "The Online Book Page" for query "search:digital library SBRank:0.5 lang:all". The first one is digital archive of past versions of pages. The second page is "Scientific Literature Digital Library". The third one lists over 20,000 free books on the Web that can be read online. We believe that these 3 results are all informative pages for query "digital library".

For the second query "search:Vancouver SBRank:0.5 time:new lang:ja" the system returned pages titled "World Peace Forum 2006@Vancouver", "Mapletown Vancouver information" and "Tourism Vancouver". The first result is a site about World Peace Forum held at Vancouver in June 2006. It contains rather fresh information related to events in Vancouver. Note that this page originally was listed in 98th search position in Google results. The other two pages are travel guides of Vancouver.

For the third query "search:wii SBRank:0.5 emo:useful lang:ja" the system returned pages titled "Wii-Tube: Let's watch YouTube with Wii", "Yahoo! News: a man confirmed effect of diet by playing wii" and "ITmedia Biz ID: Can we control PowerPoint with Wii remote?". We can see that all of these pages have different type of contents. But they are common in the sense of providing "useful" information related to wii Nintendo console.

After issuing the fourth query "search:iphone SBRank:0.5 from:20040101 to:20061201 lang:ja" the system returned pages titled "Itmedia News: Will Apple iPhone appear in next a half year?", "A Fake iPhone CM made too better: Gizmodo Japan" and "Various expectations of Apple iPhone Design – GIGAZINE". The query time span was set for the time period before iPhone was unveiled. The returned pages were thus about expectations of iPhone. This shows temporal search capabilities of the system.

Next query "search:gap-widening society SBRank:0.5 res:buzz lang:ja" produced pages titled "Japanese gap-widening society from the point of view of India, a country where the castes exist", "Daily report from mad boy - three gap-widening societies" and "A thing desired by underdog, sort of accept gap-widening society". The pages contain the content protesting against gap-widening society. They actually have many comments left by social bookmarkers.

For the sixth query "search:sns SBRank:0.5 tag:compilation lang:ja" listed pages titled "SNSLinK", "SNS list | SNS portal site SNS Navi" and "SNS Navi: SNS information portal site about SNS, building SNS, etc". These sites provide detailed and well structured information about SNS, for example, explanation of what SNS is, categorized collection of SNS links, tips for using or managing SNS, support for SNS advertise delivery, and their support for running SNS.

For the seventh query "search:web design SBRank:1 freq:cont lang:all" the returned pages were related to knowledge archive about web design. For the eighth query "search:windows vista SBRank:1 lang:ja" the top three pages returned were about Windows Vista operating system. Finally, the last query

"search:Kyoto sightseeing SBRank:1 lang:ja" produced pages about Kyoto sightseeing.

Table. 2. Example queries and their top 3 results

Query	Top 3 Results	Google
search:digital library SBRank:0.5 lang:all	Internet Archive	77th
	CiteSeer: The NEC Research Institute Scientific Literature Digital Library	19th
	The Online Books Page	2nd
search:Vancouver SBRank:0.5 time: new lang: ja	World Peace Forum 2006@ Vancouver report Blog	98th
	Mapletown Vancouver information	9th
	Tourism Vancouver	1st
search:wii SBRank:0.5 emo:useful lang: ja	Wii-Tube: Let's watch YouTube with wii	63rd
	Yahoo News: a man confirmed effect of diet by playing wii	80th
	Itmedia Biz ID: Can we control PowerPoint with Wii remote?	33rd
search:iphone SBRank:0.5 from:20040101 to:20061201 lang:ja	Itmedia News:Is Apple iPhone appear in next a half year ?	4th
	A Fake iPhone CM which made too better : Gizmodo Japan	19th
	Various expectation of Apple iPhone Design - GIGAZINE	6th
search:gap-widening society SBRank:0.5 res: buzz lang: ja	Japanese gap-widening society from the point of view of India, a country which existing the caste	23rd
	Daily report from mad boy - three gap-widening societies	5th
	A thing desired by underdog sort of accept gap-widening society	42nd
search:sns SBRank:0.5 tag:compilation lang:ja	SNSLinK	47th
	SNS list | SNS portal site "SNS Navi"	13th
	SNS Navi: SNS information portal site about SNS, building SNS, etc	12th
search:web design SBRank:1 freq:cont	Stylegala - Web Design Publication	95th
	www.welie.com -- patterns in Interaction Design	77th
	Web Design Library — One-stop resource for web designers	28th
search:windows vista SBRank:1 lang:ja	Windows Vista Encyclopedia - From install to settings, application	15th
	Irregular column by Kazuhisa Nishikawa "Some reasons I can't like Windows Vista"	11th
	FrontPage – Windows Vista Wiki	13th
search:Kyoto sightseeing SBRank:1	Let's go Kyoto – Hot Kyoto sightseeing information	46th
	For Kyoto information, e-Kyoto net – whole Kyoto portal site	5th
	Kyoto sightseeing taxi: sightseeing in Kyoto, autumn tint guide and cherry blossom information	13th

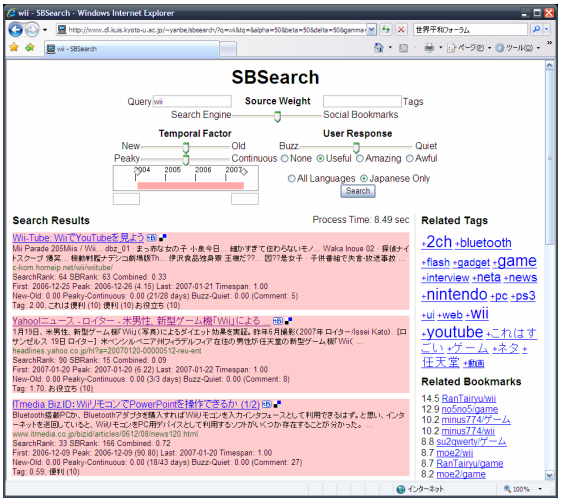

Fig. 10. Example of search result for content query "wii" with sentiment query "useful".

7. CONCLUSIONS

Social bookmarking is one of the foundations of Web 2.0. Utilizing social recommendations for complementing search in the Web appears to be quite appealing as pages are manually checked and recommended by users. These users are usually content consumers rather than content producers as in the standard link-based approach. There are many advantages of using social bookmaking data such as high dynamics, attached metadata, available temporal and sentiment information, etc.

In this paper, we investigated an enhanced search model in the Web that combines the ranking method based on link structure analysis with one based on social bookmarking. Such a search type offers many enhancements to the current search engines. First, page quality measure can be improved thanks to incorporating the popularity statistics of pages in social bookmarking systems. Second, it enables more precise relevance estimation of documents by leveraging metadata provided by users. Since bookmarks usually have timestamps provided, time-aware popularity measure is feasible and temporal queries can be constructed. Finally, tags allow for filtering of pages by user impressions, sentiment characteristics or controversy levels.

In order to support our approach, several analytical studies of social bookmarks as well as comparative analysis between PageRank and SBRank were conducted. The results allowed us to conclude that a hybrid, enhanced Web search is possible and can provide several advantages. We have implemented the prototype search system and demonstrated its search capabilities through preliminary experiments.

We believe that search systems using our approach will become popular in the future considering the current limitations of link-based ranking algorithms, the proliferation of social collaboration in the Web and the capabilities that social bookmarking systems offer. In the future, we plan to conduct larger scale experiments using different datasets and more formal evaluations. Additionally, we would like to investigate possibility of meta-search type that would aggregate data from multiple social bookmarking systems.

8. 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).

9. REFERENCES

[1] E. Amitay, D. Carmel, M. Herscovici, R. Lempel, and A. Soffer. Trend Detection Through Temporal Link Analysis. In *Journal of The American Society for Information Science and Technology*, 55, 2004, 1–12.

[2] R. Baeza-Yates, C. Castillo and F. Saint-Jean. Web Dynamics, Structure and Page Quality. In M. Levene and A. Poulovassilis (eds.) *"Web Dynamics"*, Springer, 2004, 93-109.

[3] R. Baeza-Yates, C. Hurtado, and M. Mendoza. Query recommendation using query logs in search engines. In *International Workshop on Clustering Information over the Web* (ClustWeb, in conjunction with EDBT), Creete, Greece, March, Springer, LNCS, 2004, 588-596.

[4] F. Bry and H. Wagner. *Collaborative Categorization on the Web: Approach, Prototype, and Experience Report*. Forschungsbericht/research report, 2003, 1-14.

[5] J. Cho, S. Roy, and R. Adams. Page Quality: In Search of an Unbiased Web Ranking. In *Proceedings of SIGMOD Conference* 2005, 2005, 551-562.

[6] H. Cui, J. R. Wen, J. Y. Nie, and W. Y. Ma. Probabilistic query expansion using query logs. In *Proceedings of the 11th International World Wide Web Conference*, Honolulu, Hawaii, 2002, 325-332.

[7] L. Damianos , J. Griffith, and D. Cuomo. *Onomi: Social Bookmarking on a Corporate Intranet*. The MITRE Corporation. Technical paper, 2006, www.mitre.org/work/tech_papers/tech_papers_06/06_0352.

[8] D. Fetterly, M. Manasse, M. Najork, and J.Wiener: A large-scale study of the evolution of Web pages. In *Proceedings of the 12th International World Wide Web Conference*, Budapest, Hungary, 2003, 669-678.

[9] S. A. Golder and B. A. Huberman. The Structure of Collaborative Tagging Systems, In *Journal of Information Science*, 2006.

[10] D. Gomes and M. J. Silva. Modeling information persistence on the Web. In *Proceedings of the 6th International Conference on Web Engineering*, Palo Alto, CA, USA, 2006, 193-200.

[11] W. I. Grosky, D. V. Sreenath, F. Fotouhi. Emergent Semantics and the Multimedia Semantic Web. *SIGMOD Record 31(4)*, 2002, 54-58.

[12] T. H. Haveliwala. Topic-Sensitive PageRank: A Context-Sensitive Ranking Algorithm for Web Search. In *IEEE Transactions on Knowledge and Data Engineering*, 2003, 784-796.

[13] J. M. Kleinberg. Authoritative Sources in a Hyperlinked Environment. *Journal of the ACM*, 1999, 604-632.

[14] T. Mandl. Implementation and evaluation of a quality-based search engine. In *Proceedings of ACM Hypertext 2006 Conference*, 2006, 73-84.

[15] C. Marlow, M. Naaman, D. Boyd and M. Davis. HT06, Tagging Paper, Taxonomy, Flickr, Academic Article, To Read. In *Proceedings of ACM Hypertext 2006 Conference*, Odense, Denmark, 2006, 31-40.

[16] A. Mathes. *Folksonomies - Cooperative Classification and Communication Through Shared Metadata*. Computer Mediated Communication, LIS590CMC, 2004.

[17] R. Mertens, R. Farzan, P. Brusilovsky. Social navigation in Web lectures. In *Proceedings of ACM Hypertext 2006 Conference*, Odense, Denmark, 2006, 41-44.

[18] L. Page, S. Brin, R. Motwani and T. Winograd. *The pagerank citation ranking: Bringing order to the Web*. Technical report, Stanford Digital Library Technologies Project, 1998.

[19] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing and Management: an International Journal*, 24:5, 1988, 513-523.

[20] D. N. Strutz. *Communal Categorization: The Folksonomy*. INFO622: Content Representation, 2004.

[21] X. Wu, L. Zhang and Y. Yu. Exploring Social Annotations for the Semantic Web. In *Proceedings of the 15th World Wide Web Conference*, 2006, 417-426.

[22] H. Wu, M. Zubair and K. Maly. Harvesting Social Knowledge from Folksonomies. In *Proceedings of ACM HyperText 2006 Conference*, Odense, Denmark, 2006, 111-114.

[23] L. Zhang, X. Wu and Y. Yu. Emergent Semantics from Folksonomies: A Quantitative Study. *Journal on Data Semantics VI*, LNCS 4090, 2006, 168-186.

[24] P. S. Yu, X. Li, B. Liu. On the temporal dimension of search. In *Proceedings of the 13th international World Wide Web Conference*, 2004, 448-449.