Web Search Personalization via Social Bookmarking and Tagging

Michael G. Noll and Christoph Meinel

Hasso-Plattner-Institut an der Universität Potsdam, Germany

Abstract. In this paper, we present a new approach to web search personalization based on user collaboration and sharing of information about web documents. The proposed personalization technique separates data collection and user profiling from the information system whose contents and indexed documents are being searched for, i.e. the search engines, and uses social bookmarking and tagging to re-rank web search results. It is independent of the search engine being used, so users are free to choose the one they prefer, even if their favorite search engine does not natively support personalization. We show how to design and implement such a system in practice and investigate its feasibility and usefulness with large sets of real-word data and a user study.

1 Introduction

The recent emergence and success of folksonomies and the so-called *tagging* with services such as del.icio.us or Flickr have shown the great potential of this simple yet powerful approach to collect metadata about resources. Unlike traditional categorization systems, the process of tagging is nothing more than annotating documents with a flat, unstructured list of keywords called *tags*. Although the number of peer-reviewed research on tagging is still comparatively low, several studies have already analyzed the semantic aspects of tagging and why it is so popular and successful in practice [1], [2]. A common argument is that tagging works because it strikes a balance between the individual and the community: the cost of participation is low for the individual, and tagging a document benefits both the individual and the community.

In this paper, we describe and analyze a system for personalization of web search based on such tagging metadata, i.e. user-contributed information about documents. Traditional web search has been rather impersonal: returned search results are the outcome of a function applied to the entered query. From a set of documents, those items that best match the query are returned to the user. Characteristics of the user are not taken into consideration when processing a query. Personalized web search on the other hand integrates user-specific data into the process of finding the best matching documents to a search query by increasing the amount of a priori input information available to search algorithms.

Pitkow et al. [3] describe two general approaches to web search personalization. The first modifies or augments a user's original query. For instance, a query for “nyt” might be translated to “new york times”. The second approach will run the unmodified original query for all users but re-rank the returned results based on information about the individual user. For the work in this paper, we will focus on the second case, i.e. re-ranking of the list of search results returned by a search engine. The proposed personalization technique benefits from the strategy of search engines to distribute their top search results among the various meanings and topics[1] of a query [4].

In the next section, we present our approach to web search personalization and describe its theoretical backgrounds and concepts. Section 3 outlines the technical design and implementation for running and using such a system in practice. The experiments in section 4 analyze the feasibility and usefulness of our approach with large sets of real-word data and a user study, followed by the conclusions in section 5.

2 User-driven personalization

Our approach to web search personalization is based on social bookmarking and tagging. The proposed technique for web search personalization is independent of the search engine being used, so users are free to choose the one they prefer. Personalization in this paper is defined as re-ranking the list of search results returned by a search engine [3]. Our approach exploits the conceptual links between web search, social bookmarking and tagging. From a high-level perspective, users search the Internet via a search engine, find the web documents or information they are looking for, and if the information is worth storing for later use or sharing with other users, they bookmark the web document [5]. Our approach makes use of bookmarks combined with tagging and collaboration for learning more about users *and* the web documents which are being searched for. Whenever a user bookmarks a web document, more data is available about the user and the bookmarked document and thus for personalization. We have shown in [6] that tagging metadata contains information which is not directly contained within a document, and so we argue that integrating tagging information can help to improve personalization and retrieval techniques.

From a conceptual point of view, personalization based on tagged bookmarks is a mixture of explicit and implicit personalization. Neither are users prompted to enter their preferences directly in a special configuration step, nor are they monitored or tracked in the background without being aware of it. User data is collected rather explicitly because bookmarking a web document and adding metadata like tags are manual user tasks. Users know exactly when information is collected for personalization. However, unlike traditional explicit personalization techniques, these manual tasks are not necessarily an additional burden for users. Bookmarking web documents has a direct benefit for users even without

[1] For instance, a search for “jaguar” returns links to web documents about the car, the feline and the Mac OS version of the same name in the top 10 search results in order to increase the chance that at least one of these topics matches the user’s intended search.

personalization, mitigating the practical problems of explicit feedback techniques [7]. The recent success of social bookmarking services such as del.icio.us which has a community of more than one million registered users[2], has shown that users are indeed willing to make use of bookmarking and share this information with other users. We therefore argue that expecting explicit actions from a user is reasonable as long as the "cost" and effort of the action is low compared to the user's subjective benefits and outcomes. The emergence and success of tagging (see Sect. 2.1) has been attributed to the same reasons [1], [2]. In addition, we present an easy way for automated creation of tagged bookmarks called *tagmarking* in Sect. 2.1 in order to close the usability gap to fully implicit techniques.

Our personalization technique consists of two main elements. First, the collection and aggregation of data about users and documents, and second, the personalization of web search based on this data. Normally, these steps are performed by the search engines themselves. In our approach however, information about web documents is collaboratively collected and shared by the community of users via social bookmarking, and - together with a user's individual profile based on her own bookmarks - used to personalize the generic search results returned by a search engine. This means it is possible to provide web search personalization independent of the search engine being used.

2.1 Data collection

Bookmarking When a user bookmarks a web document, she stores it for later use [5]. This observation leads to our two basic assumptions about bookmarking:

1. users only bookmark documents valuable/relevant to them (or their friends)
2. users have an incentive to add meaningful metadata to bookmarks

When a document is of no interest to a specific user, it is unlikely she would bookmark it. And when users actually do store bookmarks in order to find them again later, they have an incentive to add meaningful metadata to them, for example in the form of tags. Finding word associations for describing a document in the form of tags is a subjective user task, which should help with the differentiation of a user's characteristics when performing personalization. It is possible to help users with entering metadata, for example by presenting the list of the most frequently used tags to annotate the document [8].

"Bookmarking" a web document by a user is defined in this paper as storing a bookmark including any additional metadata at a social bookmarking service. A social bookmarking service is a central online service which enables users to add, modify, and delete bookmarks of web documents with additional metadata. The social aspect of the bookmarking service allows a user to share this information with the community. On popular social bookmarking services like del.icio.us, users can browse the bookmark collection of others and request community information about a web document (identified via its URL) from the

[2] http://blog.del.icio.us/blog/2006/09/million.html, last retrieved on July 24, 2007

service. In this paper, the prime purpose of a social bookmarking service is to collaboratively collect and share metadata about web documents in the form of tags. Use of tagging metadata implies that even such web documents may be processed for which existing content extraction and indexing techniques do not work efficiently because the depicted content is difficult to analyze (e.g., multimedia content such as videos).

Tagging The recent emergence and success of tagging with services such as del.icio.us or Flickr have shown the great potential of this simple approach to add metadata to documents. Unlike traditional classification or categorization systems, the process of tagging is nothing more than annotating documents with a flat, unstructured list of keywords called "tags". Users can browse or query documents by tags, and so-called "tag clouds", a hyperlinked collection of most frequently used tags, provide a rudimentary but often sufficient way to find popular and interesting content. Tagging can be interpreted as a relation $R_{tagging} \subseteq D \times U \times T$ where D is the set of documents, U the set of users and T the set of tags. In our case, the act of bookmarking a document with tags by a user creates one or more tuples as described by the relation above. Documents are identified by their URLs and users by their account name in the bookmarking service. We will use tags associated with bookmarks to collect information about documents and users alike.

Storing and organizing bookmarks with the help of tags mitigates some of the problems of hierarchical bookmark classification (for example, where to file a bookmark if it fits to more than one category) and increases findability. This is a benefit especially for users with lots of bookmarks, which gives yet another incentive to actively bookmark and tag web documents, which in turn improves personalization. As we will see in Sect. 4, even a modest amount of tagged bookmarks may lead to very good personalization performance.

Tagmarking In our system prototype, we have developed a browser extension which allows users to "tagmark" pages found via search queries. Tagmarking exploits the similarity between tags and search keywords [9]. The basic idea of the Tagmark extension is to store the search query, e.g. "gutenberg poe raven", in memory while the user evaluates search results. Whenever she finds a relevant web document, she can bookmark it with a single click on the "Tagmark" button, and the browser extension will automatically translate the search query to tags and add them to the bookmark (here: "gutenberg", "poe", "raven"). Tagmarking is a very convenient way to enhance the search and bookmarking experience by enabling users to store bookmarks with meaningful metadata with just a single click, and it helps to collect more input data for a user's profile and personalization. Tagmarking reduces the cost and effort of bookmarking and tagging a document, thereby closing the usability gap to personalization approaches based on fully implicit user actions.

2.2 Data aggregation

User profiling A user's bookmark collection R_u can be described as a relation $R_u \subseteq D \times T$ (cf. $R_{tagging} \subseteq D \times U \times T$ in section 2.1) and implemented as a tag-document matrix M_d with m tags and n documents (and thus n bookmarks).

$$M_d = \begin{bmatrix} c_{11} & \cdots & c_{1n} \\ \vdots & \ddots & \vdots \\ c_{m1} & \cdots & c_{mn} \end{bmatrix}, c_{ij} \in \{0, 1\}$$

A bookmark of a document d_j is a column (vector) $\boldsymbol{b_j}$ with its components c_{ij} set to 1 if tag t_i is associated with d_j and 0 otherwise. The user profile $\boldsymbol{p_u}$ is a vector with m components as follows:

$$\boldsymbol{p_u} := M_d \cdot \boldsymbol{\omega}_d = \begin{bmatrix} c_1^* \\ \vdots \\ c_m^* \end{bmatrix}, c_i^* \in \mathbb{N}_0$$

In our implementation, we define the weight $\boldsymbol{\omega_d}^T := \mathbf{1}^T = \begin{bmatrix} 1 \cdots 1 \end{bmatrix}$ with n dimensions, thereby assigning equal importance to all n documents. Here, c_i^* denotes the total count of tag t_i for the user's bookmark collection. We assume that frequently used tags are more interesting and relevant to a user than rarely used tags. Building a user's profile as described implies that it can be updated incrementally whenever a user adds a new bookmark to her collection or modifies or deletes an existing one. By this, personalization can adapt to shifts of interests over time. Table 1 shows an exemplary user profile.

User jsmith			**URL http://iswc.semanticweb.org/**	
"open source"	13		"iswc"	156
"programming"	19		"computing"	48
"proprietary"	2		"programming"	66
"research"	10		"conference"	90
"security"	21		"research"	111
"semantic web"	34		"semantic web"	140

Table 1. Exemplary profile for a user (left) and a document (right).

Document profiling Metadata about web documents is collected by the community of users submitting bookmarks to the bookmarking service. In contrast to individual user profiles, document profiles are a collaborative work. Whenever a user creates or modifies a bookmark of a web document, the information is shared with the community and the document's profile is updated.

Metadata about a document d can be described as a relation $R_d \subseteq U \times T$ and implemented as a tag-user matrix M_u with m tags and n users. A bookmark of document d by user u_j is a column (vector) $\boldsymbol{b_j}$ with its components c_{ij} set to 1 if tag t_i is associated with d by user u_j and 0 otherwise. Similar to user profiles, the document profile $\boldsymbol{p_d}$ is a vector with m components generated by $\boldsymbol{p_d} := M_u \cdot \boldsymbol{\omega}_u$. In our implementation ${\boldsymbol{\omega}_u}^T := \mathbf{1}^T = \begin{bmatrix} 1 \cdots 1 \end{bmatrix}$ with n dimensions, thereby assigning equal importance to all n users. Table 1 shows an exemplary document profile.

2.3 Personalization

After having collected data in the previous sections and transformed it into user profiles and document profiles, we can now match users and documents in the actual personalization step with the goal of re-ranking a list of search results as shown in Algorithm 1.

Algorithm 1 PERSONALIZE(*user*, *documents*)

Require: The user's profile and a sequential list of document profiles.
Ensure: The personalized list of documents for the user.

```
for all $d$ in $documents$ do
    CALCULATE $similarity(user, d)$
end for
# highest to lowest, stable sort
SORT $documents$ BY SIMILARITY
return $documents$
```

The user-document similarity is a dimensionless score used for relative weighting and re-ranking of documents within a given list, defined as:

$$similarity(u, d) := {\boldsymbol{p_u}}^T \cdot \|\boldsymbol{p_d}\|$$

where the naïve "normalization" of the document profile, $\|\boldsymbol{p_d}\|$, simply sets all non-zero components of $\boldsymbol{p_d}$ down to 1. The main idea is to leverage community-supplied metadata mostly for identifying commonly agreed "perceptions" of documents, and let the unmodified user profile be the key factor for personalization. At the moment, we are experimenting with more sophisticated normalization techniques for both user and document profiles.

The similarity for the exemplary user and the exemplary profile of the ISWC home page in Table 1 is 63. The described user-document similarity favors documents with tags that are frequently applied by the user herself, and the personalization algorithm tends to promote known, similar documents and demote non-similar or unknown documents. *Known* in this case means that documents have already been bookmarked and tagged by users. Thus, an important factor for the viability of this personalization approach in practice is the availability of user-supplied metadata for web documents, which we study in Sect. 4.1.

The left side of Table 2 shows an exemplary search query on Google for "security" by a user with a strong interest in information technology and network security. After personalization, the result list looks as shown on the right side of Table 2. Websites related to IT security have been promoted to the top, while pages such as the White House's information page about Homeland Security have been demoted to the bottom. In this example, the algorithm has confirmed the top-ranked search result of SecurityFocus for this user, so there is no change for position 1. However, the home page of CERT, a center of Internet security expertise, has been pushed from position 9 to 2. The US Department of Homeland Security lost six positions and is now at the bottom of the list. Note that the website of the US Social Security Administration has been promoted even though it is not related to IT security; this is because the user profile also shows interests in insurance matters.

#	**URL**	#	△#	**URL**
1	securityfocus.com/	1	●	securityfocus.com/
2	microsoft.com/security/	2	⇑ +7	cert.org/
3	microsoft.com/technet/security/def...	3	●	microsoft.com/technet/security/def...
4	dhs.gov/	4	⇑ +4	w3.org/Security/
5	whitehouse.gov/homeland/	5	⇑ +2	ssa.gov/
6	windowsitpro.com/WindowsSecurity/	6	⇑ +4	nsa.gov/
7	ssa.gov/	7	↓ −5	microsoft.com/security/
8	w3.org/Security/	8	↓ −2	windowsitpro.com/WindowsSecurity/
9	cert.org/	9	↓ −4	whitehouse.gov/homeland/
10	nsa.gov/	10	↓ −6	dhs.gov/

Table 2. Google search results for "security" before (left) and after personalization. URL scheme and "www." prefix omitted, long URLs have been truncated.

2.4 Putting it all together

The complete process of web search personalization works as follows:

1. the user makes a query on a search engine of her choice
2. a list of documents is returned by the search engine as result of the query
3. for each result document, the document profile is retrieved from the bookmarking service
4. the user's profile is generated on the client side (the user's computing device)
5. the list of documents is personalized based on user-document similarity

Steps 1-3 require communication between the user's client and the search engine or bookmarking service. The communication flow is shown in Fig. 1. Steps 4-5, which include the actual personalization, are performed only at the client side. The proposed personalization technique has several benefits. First, bookmarking a web document will improve future web searches even if the user is not actively using a search engine. When a user bookmarks a web document recommended to her via email, it will still affect her user profile. Second, the technique allows the personalization of search results from different search engines. Because the user is in control of her user profile at any time, i.e. it is not managed by a specific search engine, she can personalize multiple (even competing) search engines with the same user data. Third, it is even possible to personalize a search engine which natively does not support personalization of search results. Fourth, it is relatively easy to explain users why a web document has been promoted or demoted during personalization (e.g., "...because you have tagged a lot of bookmarks with..."). Fifth, the computational expense of the algorithms for generating user and document profiles and calculating user-document similarity is very low. Personalization can easily be performed on client devices with limited energy or processing power such as mobile phones or PDA.

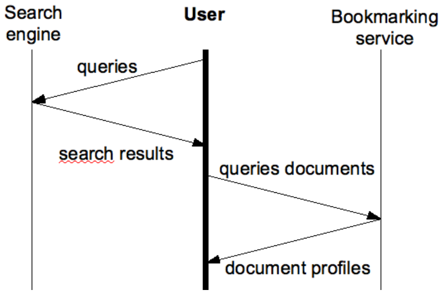

Fig. 1. Communication flow for web search personalization.

3 System setup

We have designed and implemented a system prototype for web search personalization with the following components. First, a custom social bookmarking service for adding, modifying and querying metadata about web documents, and second, a browser extension for on-the-fly personalization of search results and managing bookmarks with tags. The third required "component" is any search engine whose search results shall be personalized.

The technical implementation of the bookmarking service follows the system we have described in detail in [10]. The service allows users to add, modify and delete bookmarks of web documents with additional metadata such as tags. For each bookmarked web document, the service aggregates metadata submitted by the user community into a document profile, which can be accessed via a web API. The browser extension is installed on the user's computing device. The extension is responsible for carrying out steps 3-5 as described in Sect. 2.3 by communicating with the bookmarking service over its web API. It will transparently personalize the search result pages of search engines like Yahoo or Google by modifying the DOM tree of these web pages on-the-fly. It will also highlight web documents already bookmarked by the user for easier reference[3]. From a user perspective, web search personalization in our system setup is completely transparent and happens instantly even though extra communication with the bookmarking service is required. The technical implementation of DOM tree manipulation, i.e. displaying the personalization results to the user via the browser UI, is specific to a particular search engine. On a conceptual level, however, the personalization of search results is independent of the search engine being used. The extension enhances the browser GUI with interface elements for saving tagged bookmarks to the bookmarking service and features a “tagmarking” button as described in Sect. 2.1.

4 Experiments and evaluation

Personalization in this paper relies on the strength of the user community as it requires that search result documents have been tagged by users. For documents without bookmarks or tags, our personalization approach is not possible in practice because metadata about them is missing and thus document profiles cannot be generated. One if not the most important task is therefore to analyze the expected availability of metadata for search result documents in the real-world. However, the custom bookmarking service we have implemented limits the possibilities of sharing and comparison of research results. We have therefore decided to use the public bookmarking service of del.icio.us with its large user community (more than one million registered members in 2006) as information source for our experiments.

4.1 Quantitative analysis

In a previous study [6], we analyzed the availability of user-contributed metadata for a random sample of 100,000 web documents from the Open Directory. One finding was a correlation between tagging metadata and a document’s popularity (measured by its Google PageRank): the more popular a web page, the more likely the page is to be bookmarked and tagged by users. We can thus infer information about bookmarking and tagging metadata of search result documents by analyzing their PageRank distribution. For the work in this paper, we combined our previous results with an analysis of the AOL500k corpus[4], of which we evaluated ∼1,750,000 queries with 1,000,000 clicked search results. For each clicked document, we retrieved PageRank information from Google.

[3] We are currently working on a feature that will retrieve potentially relevant bookmarks from a user’s bookmark collection based on the entered search query. This will allow us to present search results to the end user which are not in the search engine’s index at all, e.g. a bookmark to a non-public intranet web page.

[4] The AOL500k corpus is a collection of ∼20,000,000 search queries from ∼650,000 anonymized but uniquely identifiable users sampled by AOL over a period of three months in 2006. The corpus was formerly available from http://research.aol.com/.

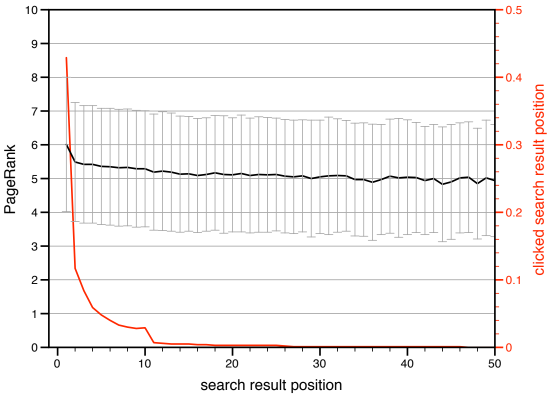

Fig. 2. Average PageRank of result links (black) with error bars (gray) denoting standard deviations. Click frequency of users is shown by the spiked red line.

First, we looked at documents and analyzed the average PageRank of web pages for each search result position. The top positions have an average PageRank of 5.4 or higher as shown by the black line in Fig. 2. The red line denotes the click frequency per search result position. The top 5 positions account for approx. 75% of all clicked search results, most of which is contributed by position 1. The drop between position 10 and 11 is caused by the default configuration of AOL search to show only 10 results per result page (similar to most popular search engines), which means that users are very unlikely to look for search results beyond page 1. This result is encouraging for our personalization technique. On one hand, search results documents are likely to be bookmarked and tagged due to the high expected PageRank, and on the other hand, the re-ranking approach can prove to be efficient in practice because it is very often sufficient to personalize just the first result page.

Second, we looked at users and averaged the PageRank of clicked search results for each user in our AOL500k subset, i.e. individual click preferences regardless of search result position or result page. The black line in Fig. 3 shows the percentage of users with an average clicked PageRank of x or higher. 80% of users have an average clicked PageRank of 5 or higher, 33% a PageRank of 6 or higher. The dashed and dashed+dotted lines describe the probability of a document to be bookmarked or tagged, respectively, based on our findings in [6]. While the numbers for PageRanks less than 5 might seem low at first glance, the *combined* probability of n result documents with varying PageRank to be bookmarked or tagged can be high enough in practice for good personalization results as we will see later. Additionally, the usage of social bookmarking services and collaboration platforms such as del.icio.us, on which the evaluations in [6] is based, is increasing in the Web today, and thus the availability of tagging metadata will increase over time, too.

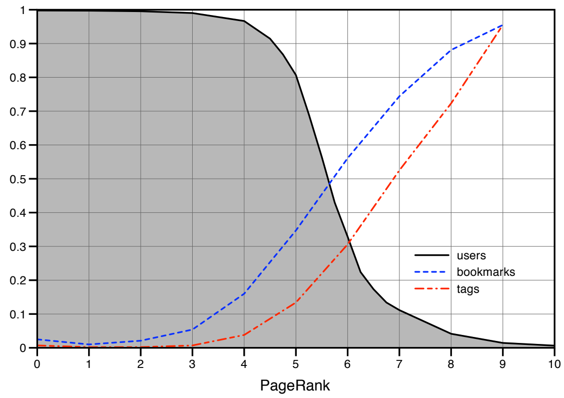

Fig. 3. Percentage of users with an average PageRank for clicked search results equal to or higher than x, i.e. $P_u(PageRank \geq x)$. The dashed and dashed+dotted lines denote the frequencies of bookmarked or tagged documents for PageRank x, respectively, i.e. $P_d(bookmarked|PageRank = x)$ and $P_d(tagged|PageRank = x)$.

In the next experiment, we extracted the so-called *popular tags* from del.icio.us and run searches for each tag on Google. For each document on the first result page (see above), we retrieved the document's *common tags*[5] from del.icio.us, similar to steps 1-3 in Sect. 2.4. The final data set consisted of 140 tags and associated search queries with 1400 result links. A total of 981,989 user bookmarks with 20,498 tags (2,300 unique) were stored at del.icio.us for these 1400 web documents, netting 701 bookmarks and 14.6 tags per document in average. The full details are shown in Table 3. Around 9 out of 10 search results are bookmarked and 8.5 out of 10 search results are tagged by users (see Fig. 4). In other words, we can expect to be able to personalize approx. 85% of search results per query in practice - at least for popular keywords - when using the del.icio.us bookmarking service as the sole information source for tagging metadata.

4.2 Qualitative analysis

To examine the usefulness of our personalized search system for individual users, we let 8 participants evaluate the top 10 search results, i.e. the first result page, for 13 queries each, totaling 104 queries. Each user had her or his personal set of bookmarks, which was used to build the individual user profile. The public bookmark repository of del.icio.us was used for creating the document profiles.

[5] del.icio.us limits a document's *common tags* to its 25 most popular tags, which means that the list of all tags attached to a document might actually be larger in practice. The reason for retrieving just the common tags of a document instead of all tags, i.e. even rarely used ones, is due to technical restrictions by del.icio.us. Still, we argue that even if all tagging information was available, it would be recommended to perform some sort of thresholding or preprocessing anyway to remove "tag noise".

#	Bookmarks	Tags		#	Bookmarks	Tags
1	1450	19.8		6	456	13.7
2	627	16.4		7	495	13.4
3	1199	15.5		8	574	13.7
4	451	14.2		9	404	14.0
5	610	12.5		10	784	13.3

Table 3. Average number of bookmarks and tags of a document per search result position. The peak of 784 for position 10 is caused by two extreme data points in our sample; it drops to 519 when these two data points are removed.

Search queries varied by user by their bookmarking history. Web search results were collected from Google Search. For each query, participants were presented two search result lists: the original, "generic" list from Google Search, and the personalized version. The experiment was conducted as a blind test, i.e. the result lists were presented in random order so as not to bias the participants. Participants were asked to determine which of the two results lists of a query was "better" tailored to them, where *better* was defined as ranking highly relevant results at the top of the lists and ranking irrelevant results at the bottom, i.e. promoting "good" results and demoting "bad" results. Participants could also vote for a draw if they could not decide which list was better. The participant's job functions included researchers, system administrators, webmasters and software developers. All were computer literate and familiar with web search. The average number of bookmarks for a participant was 153.

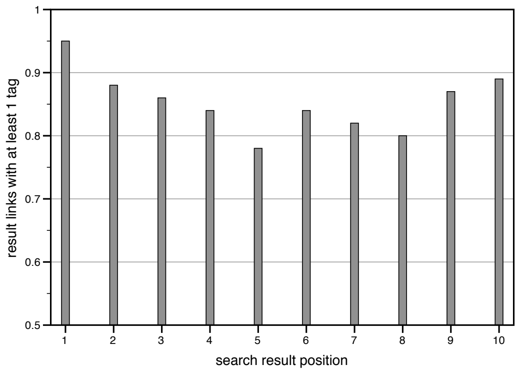

Fig. 4. Percentage of search result links which have at least 1 associated del.icio.us "common" tag.

In our experiment, the personalized list was considered better than or as good as the unmodified result list in 70.2% of the queries (63.5% and 6.7%, respectively). The unmodified result list as returned by the search engine was preferred in 29.8% of the queries. An interesting observation was the low frequency (6.7%) of the cases where users could not prefer one list over the other. The personalization had problems to improve search results for users which were only broadly interested in a particular topic, suggesting that it performs better for "expert" user profiles. For instance, a user with lots of bookmarks tagged just with "web design" will not benefit as much from personalization as a user who tags her bookmarks about web design more granularly. Previous studies such as [11] have shown that most search queries are rather short, with the average search query consisting of only one or two words. Our approach showed its strength particularly in the case of disambiguation of words and contexts (see the "security" example in Sect. 2.3), which indicates that it is especially helpful for such queries.

5 Conclusions

In this paper, we presented a new approach to personalization of web search by leveraging social bookmarking and tagging. We have shown how to design and implement such a system in practice and investigated its feasibility and usefulness with large sets of real-word data and a user study. Our evaluation results are encouraging and suggest that personalization based on social bookmarking and tagging is a useful addition to the web toolset and that the subject is worth further research, in particular with regard to the increasing popularity of social and collaborative services in the WWW today.

6 Related work

To the best of our knowledge, this is the first study on using social bookmarking and tagging techniques for personalization of web search and its evaluation in a real-world scenario. Next to the references mentioned throughout the text, the following works are related to the work described in this paper. Bao et al. [12] use social annotations to improve page ranking in generic web search. They propose a similarity measure between social annotations and web queries, and use tagging information to measure the popularity of a web page from an end user's perspective. Next to the different research focus, an important difference to our work is that their experimental data set is restricted to web pages already stored at del.icio.us whereas our evaluation is based on a indiscriminate, random sample of web pages. Boydell and Smyth [9] describe a technique for document summarization that uses informational cues as the basis for summary construction. Social bookmarks are one of the cues used in their work, and they stress the similarity between tags and search query keywords for creating snippet texts for summarization. Integration of social bookmarking information helped to improve the quality of their system when compared with traditional summarization techniques. Sugiyama et al. [13] integrate collaborative filtering techniques into search personalization similar to the social bookmarking approach in this paper. However, the collaborative aspects focus on identifying similar users based on their daily browsing histories, not on sharing information about the documents being searched for as is the case for social bookmarking. In addition, the input data required for their collaborative filtering algorithms, i.e. detailed browsing history information about other users, is generally not available to an individual user of a search engine. Similar to tagging information supplied by end users, ranking and classification techniques may use incoming or outgoing hyperlinks of a web document to infer information about the document and its neighbors by associating terms with the web documents that are themselves not part of the documents [14]. Here, the descriptive annotations of other document authors (as opposed to the document readers in the case of social bookmarking and tagging) help to gain more knowledge about the documents at hand.

References

1. Golder, S.A., Huberman, B.A.: Usage patterns of collaborative tagging systems. J. Inf. Sci. **32**(2) (2006) 198–208
2. Marlow, C., Naaman, M., Boyd, D., Davis, M.: Ht06, tagging paper, taxonomy, flickr, academic article, to read. In: Proceedings of HT '06. (2006) 31–40
3. Pitkow, J., Schütze, H., Cass, T., Cooley, R., Turnbull, D., Edmonds, A., Adar, E., Breuel, T.: Personalized search. Commun. ACM **45**(9) (2002) 50–55
4. Wedig, S., Madani, O.: A large-scale analysis of query logs for assessing personalization opportunities. In: Proceedings of SIGKDD '06. (2006) 742–747
5. Abrams, D., Baecker, R., Chignell, M.: Information archiving with bookmarks: personal web space construction and organization. In: Proc. of SIGCHI '98. (1998) 41–48
6. Noll, M.G., Meinel, C.: Authors vs. readers: A comparative study of document metadata and content in the www (to appear). In: Proc. of ACM DocEng '07. (2007)
7. Carroll, J.M., Rosson, M.B.: Paradox of the active user. In Carroll, J.M., ed.: Interfacing Thought: Cognitive Aspects of Human-Computer Interaction. Bradford Books (1987) 80–111
8. Sen, S., Lam, S.K., Rashid, A.M., Cosley, D., Frankowski, D., Osterhouse, J., Harper, F.M., Riedl, J.: tagging, communities, vocabulary, evolution. In: Proc. of CSCW '06. (2006) 181–190
9. Boydell, O., Smyth, B.: From social bookmarking to social summarization: an experiment in community-based summary generation. In: IUI '07: Proceedings of the 12th international conference on Intelligent user interfaces. (2007) 42–51
10. Noll, M.G., Meinel, C.: Design and anatomy of a social web filtering service. In: Proceedings of the CIC '06, Hong Kong (2006) 35–44
11. Jansen, B.J., Pooch, U.: A review of web searching studies and a framework for future research. J. Am. Soc. Inf. Sci. Technol. **52**(3) (2001) 235–246
12. Bao, S., Xue, G., Wu, X., Yu, Y., Fei, B., Su, Z.: Optimizing web search using social annotations. In: Proceedings of WWW '07. (2007) 501–510
13. Sugiyama, K., Hatano, K., Yoshikawa, M.: Adaptive web search based on user profile constructed without any effort from users. In: Proc. of WWW '04. (2004) 675–684
14. Brin, S., Page, L.: The anatomy of a large-scale hypertextual web search engine. In: Proceedings of WWW7. (1998) 107–117