Personalizing Web Search with Folksonomy-based User and Document Profiles

David Vallet[1,2], Iván Cantador[1,2], Joemon M. Jose[1]

[1] University of Glasgow, Glasgow, UK
{dvallet, cantador, jj}@dcs.gla.ac.uk

[2] Universidad Autónoma de Madrid, Madrid, Spain
{david.vallet, ivan.cantador}@uam.es

Abstract. Web search personalization aims to adapt search results to a user based on his tastes, interests and needs. The way in which such personal preferences are captured, modeled and exploited distinguishes the different personalization strategies. In this paper, we propose to represent a user profile in terms of social tags, manually provided by users in folksonomy systems to describe, categorize and organize items of interest, and investigate a number of novel techniques that exploit the users' social tags to re-rank results obtained with a Web search engine. An evaluation conducted with a dataset from Delicious social bookmarking system shows that our personalization techniques clearly outperform state of the art approaches.

1 Introduction

The huge and ever increasing volume and complexity of information available in the Web constitutes a difficult challenge for content retrieval technologies. In a traditional Web search system, such as Google[1] or Yahoo![2], a user expresses his information needs by providing a textual query consisting in a limited number of keywords. The search system takes as input this query, and attempts to retrieve the Web documents that best match its keywords. Queries are usually short – containing no more than 3 keywords on 85% of the times – and ambiguous [7], and often fail to reflect the user's needs. Nonetheless, although the information contained in these keywords rarely suffices for the exact determination of the user's wishes, this approach represents a simple way of interaction users are accustomed to. There is thus a need to investigate ways to enhance information retrieval, without altering the way the users specify their requests. It is in such scenario where personalized information retrieval techniques can help the users, by tailoring the search results based on both the users' short and long term preferences [4]. However, to achieve that goal, information about the users' tastes and interests has to be found in other sources.

With the advent of the Web 2.0, social tagging systems have been exponentially grown both in terms of users and contents. These systems allow the users to provide annotations (*tags*) to resources, expressing personal descriptions and opinions about

[1] Google, http://www.google.com/
[2] Yahoo! Search, http://search.yahoo.com/

the resources for organizational and sharing purposes. For instance, in Last.fm[3], the users annotate their favorite songs; in Flickr[4], the users store and annotate their own photo streams; and in Delicious[5], the users bookmark and annotate interesting Web pages. Apart from facilitating the organization and sharing of content, these 'social tagging' data, also known as *folksonomies*, can be considered as a fairly accurate source of user interests. Several studies have proven that a user profile can be effectively harvested from these systems [1, 11], and later exploited on different personalization services, such as tag recommendation [3], item recommendation [9], and personalized search [6, 9, 13], to name a few.

In this work, we present two novel personalization techniques that exploit a user profile defined within a social tagging system to re-rank the document lists retrieved by a traditional Web search engine. In particular, we investigate whether a folksonomy-based user profile defined in Delicious social bookmarking system can really enhance the results provided by Yahoo! Search engine. To evaluate such techniques, we propose an automatic mechanism that generates test datasets from social tagging corpora. The results obtained in our experiments show that our personalization techniques clearly outperform state of the art approaches.

The rest of the paper has the following structure. In Section 2, we describe works that are related to our research. In Section 3, we present the state of the art and own personalization approaches we evaluate and compare. In Section 4, we propose an evaluation framework and an experimental methodology for folksonomy-based Web search personalization techniques. We present the evaluation results in Section 5. Finally, in Section 6, we provide some conclusions and possible future work lines.

2 Related Work

Personalized retrieval models that exploit user profiles based on social tags have been investigated in previous works.

Shepitsen et al. [9] present a strategy that clusters the entire space of tags to obtain sets of (semantically) related tags. Representing coherent topic areas, the obtained clusters are used to provide personalized item recommendations. Rather than item recommendation, the techniques presented in this paper follow personalized retrieval models applicable to Web search, where lists of search results are re-ranked according to the user's preferences.

Hotho et al. propose the FolkRank algorithm [4], an adaptation of the PageRank algorithm to the folksonomy structure. FolkRank performs a weight-spreading ranking scheme on folksonomies. It transforms the hypergraph between the sets of users, tags and resources into an undirected, weighted, tripartite graph. On this graph, it applies a version of PageRank that takes into account the obtained edge weights. Among other applications, FolkRank provides a popularity measure of a document that seems to be better than PageRank, as it exploits the user generated folksonomy,

[3] Last.fm - Personal online radio, http://www.last.fm/
[4] Flickr - Photo sharing, http://www.flickr.com/
[5] Delicious - Social bookmarking, http://delicious.com/

rather than the Web links. Bao et al. [2] also investigate the use of popularity measures derived from the folksonomy structure, but focusing its application in a Web search system. They introduce two importance score values, SocialSimRank and SocialPageRank, which measure the relevance of a document to a query, and the popularity of a document, respectively. They conclude that these measures provide a better performance than traditional measures, such as term matching and PageRank. Similar to the studies of Hotho et al. and Bao et al., we exploit the folksonomy structure, but focus on offering a personalized search to the user, rather than improving the overall rank of documents.

Noll and Meinel [6] present a personalization model that exploits the user's and documents' related tags, improving a Web search system during their user evaluation. Xu et al. [13] also present a user-document similarity function that relates the user's and documents' tags, and enrich the user's profile representation following a tag expansion strategy, which is applied over a restricted corpus. Our personalization approaches follow the same personalization model as Xu et al.'s and Null and Meinel's, but utilize different techniques to calculate the user-document similarities. We shall evaluate and compare our proposals against the approaches presented by these authors.

3 Web Search Personalization based on Folksonomies

A folksonomy $\mathcal{F}$ can be defined as a tuple $\mathcal{F} = \{\mathcal{T}, \mathcal{U}, \mathcal{D}, \mathcal{A}\}$, where $\mathcal{T} = \{t_1, \dots, t_L\}$ is the set of tags that comprise the vocabulary expressed by the folksonomy, $\mathcal{U} = \{u_1, \dots, u_M\}$ and $\mathcal{D} = \{d_1, \dots, d_N\}$ are respectively the set of users and the set of documents[6] that annotate and are annotated with the tags of $\mathcal{T}$, and $\mathcal{A} = \{(u_m, t_l, d_n)\} \in \mathcal{U} \times \mathcal{T} \times \mathcal{D}$ is the set of assignments (annotations) of each tag t_l to a document d_n by a user u_m.

The profile of user u_m is then defined as a vector $\overrightarrow{u_m} = (u_{m,1}, \dots, u_{m,L})$, where $u_{m,l} = |\{(u_m, t_l, d) \in \mathcal{A} | d \in \mathcal{D}\}|$ is the number of times the user has annotated documents with tag t_l. Similarly, the profile of document d_n is defined as a vector $\overrightarrow{d_n} = (d_{n,1}, \dots, d_{n,L})$, where $d_{n,l} = |\{(u, t_l, d_n) \in \mathcal{A} | u \in \mathcal{U}\}|$ is the number of times the document has been annotated with tag t_l. In our Web search scenario, the set of documents $\mathcal{D}$ represents the resources present in the Web, and are identified by an URL. Users are identified by a user id.

In this work, we exploit folksonomy-based user and document profiles in order to personalize the results of a Web search system. A non-personalized Web search system S provides a ranked list of documents $S(q) \subseteq \mathcal{D}$ that satisfy a given query topic q. The ranking follows an ordering $\tau = [d_1 \geq d_2 \geq \cdots \geq d_k]$, in which $d_i \in \mathcal{D}$ and $\geq$ is the ordering relation implemented by the search system. Upon this ranked document list, we define a personalization approach S' that provides a ranked list of documents $S'(q, u) \subseteq \mathcal{D}$ by reordering the results $S(q)$ according to the preferences of

[6] In more general definitions of *folksonomy*, annotated items, which are not necessarily textual, are usually called "resources". As we deal with the exploitation of folksonomies in a Web Search scenario, we instead use the term (Web) "documents" to reference resources.

user u. More formally, it provides an ordering $\tau' = [d_1 \geq d_2 \geq \cdots \geq d_k]$ such that the ordering relation is defined by $d_i \geq d_j \Leftrightarrow \mathrm{sim}(u, d_i, q) \geq \mathrm{sim}(u, d_j, q)$, where $\mathrm{sim}(u, d, q)$ is a similarity function between user u and document d, taking into consideration the ranking of d in $\mathrm{S}(q)$.

The subsequent subsections present the different personalization techniques we propose and evaluate. The first two techniques are obtained from the state of the art, and are based on the Vector Space Model (VSM). The third technique is a personal adaptation of the VSM to social tagging profiles. The last technique is a novel personalization approach that follows a probabilistic model. More specifically, it is an adaptation of the Okapi BM25 ranking model.

For a better understanding, Table 1 gathers the definition of common elements appearing in the models of the above techniques. It is worth noting that whereas in the classic VSM the document collection is the only source for the calculation of term frequencies and inverse document frequencies, in a folksonomy-based framework, we can also consider how informative the tags (terms) are in the user set. Thus, the user-based tag frequency $tf_{u_m}(t_l)$ measures how relevant a tag t_l is to a user u_m, and the user-based tag inverse frequency $iuf(t_l)$ measures how common or popular a tag t_l is across all users $\mathcal{U}$. The presented approaches can also be differentiated by how these local and global importance values are exploited.

Table 1. Elements that are used by the folksonomy-based personalization models

Element	Definition
User-based tag frequency	$tf_{u_m}(t_l) = u_{m,l}$
Document-based tag frequency	$tf_{d_n}(t_l) = d_{n,l}$
User-based inverse tag frequency	$iuf(t_l) = \log\frac{M}{n_u(t_l)}, n_u(t_l) = \left\lvert\{u_m \in \mathcal{U} \mid u_{m,l} > 0\}\right\rvert$
Document-based inverse tag frequency	$idf(t_l) = \log\frac{N}{n_d(t_l)},\ n_d(t_l) = \left\lvert\{d_n \in \mathcal{D} \mid d_{n,l} > 0\}\right\rvert$
User size	$\lvert u_m\rvert = \sum_{l=1}^{L} u_{m,l}$
Document size	$\lvert d_n\rvert = \sum_{l=1}^{L} d_{n,l}$

3.1 Cosine Similarity based Personalization

The approach presented by Xu et al. [13] uses the classic cosine similarity measure to compute the similarity between user and document profiles. As weighting scheme, it uses $tf\text{-}idf$[7]. Following our notation, their approach can be defined as follows:

$$\cos_{tf\text{-}idf}(u_m, d_n) = \frac{\sum_l \left(tf_{u_m}(t_l) \cdot iuf(t_l) \cdot tf_{d_n}(t_l) \cdot idf(t_l) \right)}{\sqrt{\sum_l (tf_{u_m}(t_l) \cdot iuf(t_l))^2} \cdot \sqrt{\sum_l (tf_{d_n}(t_l) \cdot idf(t_l))^2}},$$

[7] Xu et al. do not specify if they use the user-based or the document-based inverse tag frequency weights, or both. We chose to use both, as it gave the best performance values.

where the numerator is the dot product of the $tf\text{-}iuf$ and $tf\text{-}idf$ vectors associated with the user and the document, respectively. The denominator is the user and document length normalization factors, calculated as the magnitude value of those vectors. Xu et al. use a weighting scheme based on the BM25 model, this variation will be henceforth denoted as $\cos_{bm25}(u_m, d_n)$. See Section 3.4 for more details on this approach.

3.2 Scalar Tag Frequency based Personalization

The approach presented by Noll and Meinel [6] is similar to the cosine-based approach, but does not make use of the user and document length normalization factors, and only uses the user tag frequency values. The authors normalize all document tag frequencies to 1, since they want to give more importance to the user profile when computing the similarity measures. Following the notation given in Table 1, their similarity measure can be defined as follows:

$$tf(u_m, d_n) = \sum_{l:d_{n,l}>0} tf_{u_m}(t_l) \ .$$

3.3 Scalar $tf\text{-}if$ based Personalization

Next, we present our first proposed personalization approach. Similarly to Xu et al. [13], we use the $tf\text{-}idf$ weighting scheme. We eliminate however the user and document length normalization factors. In the classic VSM, the finality of the length normalization factor is to penalize the score of documents that contain a high amount of information, and might have matched the query only by chance. In terms of a social tagging system, a high amount of related tags is correlated with the popularity of the documents among users. Hence, if we used a length normalization factor, we would penalize the score of popular documents. As several works point out, this popularity value is a good source of relevancy [2, 4]. Thus, it would not be advisable to penalize popular documents. Note that eliminating the user length normalization factor does not have any effect, as it is constant in all user-document similarity calculations.

The main difference between our approach and Noll and Meinel's [6] is that we incorporate both the user and document tag distribution global importance factors, i.e. iuf and idf, following the VSM principle that as more rare a tag is, the more important it is when describing either a user's interests or a document's content. We do not normalize the content of the documents, as we believe that the distribution of tags on a document may give insights on how important a tag is to describe its content. This personalization approach can thus be defined as following:

$$tf\text{-}if(u_m, d_n) = \sum_l \left(tf_{u_m}(t_l) \cdot iuf(t_l) \cdot tf_{d_n}(t_l) \cdot idf(t_l) \right) .$$

3.4 BM25 based Personalization

The novel personalization approach presented in this section differs from the previously presented ones in that it follows a probabilistic model, rather than the classic VSM. We adapt the Okapi BM25 ranking model [8] to a personalization ranking of similarity between a user and a document. The BM25 model computes a ranking

score function of a document given a query. We can then adapt this model in two different ways: 1) by assuming that the user profile takes part as a query indicating the user's interests, or 2) by assuming that the document takes part as a query, and is matched against all user profiles. The former option will be henceforth denoted as $bm25_{u_m}$ and the latter as $bm25_{d_n}$. We first define both score functions for a single tag t_l:

$$bm25_{u_m}(t_l) = iuf(t_l) \cdot \frac{u_{m,l} \cdot (k_1+1)}{u_{m,l} + k_1\left(1-b+b \cdot {|u_m|}/{avg(|u_m|)}\right)},$$

$$bm25_{d_n}(t_l) = idf(t_l) \cdot \frac{d_{n,l} \cdot (k_1+1)}{d_{n,l} + k_1\left(1-b+b \cdot {|d_n|}/{avg(|d_n|)}\right)},$$

where b and k_1 are set to the standard values of 0.75 and 2, respectively. Then, we define the two variations of this personalization approach:

$$bm25_{u_m}(d_n, u_m) = \sum_{(l|d_{n,l}>0)} bm25_{u_m}(t_l),$$

$$bm25_{d_n}(d_n, u_m) = \sum_{(l|u_{m,l}>0)} bm25_{d_n}(t_l)\ .$$

Xu et al. [12] compute a VSM based cosine similarity measure with a weighting scheme inspired by the BM25 retrieval model. Following the notation of this section, this measure can be defined as follows:

$$\cos_{bm25}(d_n, u_m) = \frac{\sum_l \left(bm25_{u_m}(t_l) \cdot bm25_{d_n}(t_l)\right)}{\sqrt{\sum_l (bm25_{u_m}(t_l)^2)} \cdot \sqrt{\sum_l (bm25_{d_n}(t_l)^2)}}$$ [8]

4 Evaluating Folksonomy-based Personalization Approaches

Noll and Meinel [6] evaluated their personalization approach combined with a Web search engine. They adopted a user centered evaluation by creating a set of predefined queries, and by asking users to evaluate the results. More specifically, users were asked to evaluate which result list they preferred: either the Web search ranking or the personalized ranking. Xu et al. [13] used the social bookmarking information to create an automatic evaluation framework. The main advantage of their framework is that the experiments could be reproduced. However, they did not explore the performance of their personalization approaches when combined with a Web search engine. They combined their approach with a search system that was limited to the bookmarks pertinent to their test beds, ranging from 1K to 15K Web documents. The goal of our evaluation framework falls in the middle of these two approaches: 1) as Noll and Meinel, we are more interested in testing our personalization approach in a real Web search environment; and 2) as Xu et al., we adopt an automatic evaluation framework with a test bed of topics and relevance judgments extracted from the social bookmarking information. In this section, we describe our evaluation framework, highlighting the main differences between it and the previously presented.

[8] Xu et al. use a slightly modified version of the *idf* measure: $\log((\mathrm{M} - n(t_l) + 0.5)/(n(t_l) + 0.5))$, using $n_u(t_l)$ and $n_u(t_l)$ on the $bm25_{u_m}(t_l)$ and $bm25_{d_n}(t_l)$, respectively. The reported results make use of this measure.

4.1 Topic and Relevance Judgment generation

We split the tagging information of a given user into two parts. The first part forms the user profiling information, whereas the second is used for the automatic topic generation process. Hence, the subset of tag assignments used in the topic generation process is not included in the user profile, constitutes our test dataset, and thus is not part of our training dataset. This splitting process is applied to all users belonging to the initial test bed collection. Figure 1 outlines how the partition is made.

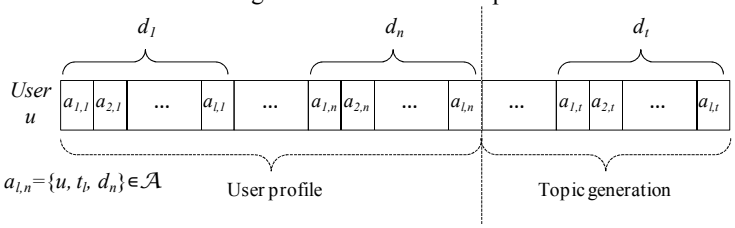

Fig. 1. Partitioning of user tag assignments into user profile and information intended for topic generation

As shown in the figure, the topic creation process attempts to create a new topic from each document $d \in [d_{n+1}, \ldots, d_t]$ belonging to the test split part. A topic is defined by extracting the top most popular tags related to a document d. We use the most popular tags as they are more objective to describe the document contents than those assigned by a single or few users. These tags are used to launch a Web search, and we collect the retrieved result list.

We then study how the different personalization approaches re-rank the returned result list. As document d was contained in the original user profile, we can assume that the document is relevant to the user. Thus, a good personalization approach would always rank the document in the top positions of the result list.

We use the Mean Reciprocal Rank (MRR) [12] metric to measure the performance of the personalization techniques. This measure assigns a value of performance for a topic of $1/r$, where r is the position of the relevant d in the final personalized result list. We also provide the *P@N* (Precision at position *N*) metric, which has a value of 1 iff $r \leq N$. These values are averaged over all the generated topics.

The topic generation and evaluation can be summarized in the following methodology. For each document $d \in [d_{n+1}, \ldots, d_t]$: 1) we generate a topic description using the top k most popular tags associated to the document; 2) we execute the topic on a Web search system and return the top R documents as the topic's result list; 3) if document d is not found in the result list, we discard the topic for evaluation; 4) we apply the different personalization techniques to the result set; and 5) we compute MRR and P@N values.

In our experiments, we used a query size of $k = 3$ tags, and a result list size of $R = 300$ documents. Several studies point out an average user query size of 2-3 keywords in Web search [7]. We thus opted for a query size of 3 in order to emulate a user using a Web search system, and to evaluate if user profiles obtained from the social tagging actions of the users could be successfully exploited to improve a Web search system. We also investigated the generation of query topics with 2 keywords

obtaining performance results similar to those obtained with topic sizes of 3 keywords. There is of course a chance that document d does not appear in the result list. In this case, the document is discarded for topic generation. With these settings, 24.2% of the topics were successfully generated, and the average position of document d on the result list was 62.2.

As mentioned before, Xu et al. also presented an automatic topic generation methodology based on the users' tagging data. However, there are some key differences between their evaluation framework and ours. First, they applied the personalization techniques to a custom search engine that only retrieves documents that belong to the same test bed. Our methodology, on the other hand, makes use of Web search system to return the topic document. In this way, we intend to have a more realistic set up. Second, they used each tag of the user profiles as a query topic, thus resulting on queries with a single keyword. This resulted on too broad queries, which are not suitable for a free Web search system. We rather choose to use more specific queries of three keywords, which are generated based on the social tagging information associated to a document that was originally in the user profile. Third, their approach assumed that a returned document was relevant to the user if it was tagged by him with the same tag that belonged to the topic query. Our ground truth is more restrictive, as we only consider as relevant the document that generated the topic query. By doing this we can ensure that the document is relevant to both the topic query, as the query keywords represent the people's view of the document's content, and to the user, as the document belongs to the user's profile. In summary, we consider that our approach is more suited to evaluate folksonomy-based personalization of a Web search system. Nonetheless, we do believe that both approaches may complement each other in order to give more insights on the performance of personalization strategies.

4.2 Experimental Setup

We created a test bed formed by 2,000 Delicious users. Delicious is a social bookmarking site for Web pages. As of the 26th of November of 2008, delicious had 5.3 million users[9], up from 1 million users registered on September of 2006[10]. With over 180 million unique URLs, Delicious can be considered a fairly accurate "people's view" of the Web. This vast amount of user information has been previously successfully exploited to improve Web search [2], to provide personal recommendations [4, 9], and to personalize search results [6, 13], among others.

Due to limitations of Delicious API, we only extracted the latest 100 bookmarks of each user, from which we use 90% of the bookmarks to create the user profile, and the remaining 10% to generate the evaluation topics as described in Section 4.1. The test bed contained 161,542 documents and 69,930 distinct tags. We did not apply any pre-processing steps to the user tags. Users used an average of 5.6 tags to describe each bookmark. As experimental Web search system, we use Yahoo!'s open Web

[9] http://blog.delicious.com/blog/2008/11/delicious-is-5.html
[10] http://blog.delicious.com/blog/2006/09/million.html

search platform, Yahoo! Boss[11]. After the topic generation process, we ended up with 6,109 evaluation topics. For each document in the topic result set, we downloaded the 100 most recent bookmarks. Those documents had an average of 24.3 distinct associated tags. On average, 20.13% of the documents of the result list had been bookmarked at least once by a user.

5 Experiment Results

We present the results of the proposed personalization techniques within the evaluation framework explained in Section 4. We first provide the performance of the approaches when applied in isolation to the search results returned by the Web search system. Then, we show their performance when taking into consideration the result ranking provided by the Web search system.

5.1 Results of Personalization Approaches

In this section, we analyze the performance of the personalization approaches when only the personalization scores are used to reorder the results returned by the Web search system, i.e., when the ranking given by the search system is not taken into account. Table 2 shows Mean Reciprocal Ranking (MRR) and Precision (at 5, at 10, at 20) values of the personalization approaches.

Table 2. Personalization approaches performance. Values with an asterisk indicate a statistically significant higher value than the tf approach (Wilcoxon test, $p < 0.05$). Values in bold indicate the highest values with statistical significance. The column $comb$ refers to the rank-based combination of $bm25_{u_m}$ and $tf\text{-}if$ approaches

Metric	$\cos_{tf\text{-}idf}$ [13]	$\cos_{bm25}$ [13]	$bm25_{d_n}$	tf [6]	$bm25_{u_m}$	$tf\text{-}if$	$comb$
MRR	0.0809	0.0912	0.2878	0.2845	0.3055*	0.3084*	**0.3241***
P@5	0.0915	0.1111	0.4502	0.4554	0.4601	0.4839*	**0.4924***
P@10	0.1838	0.2252	0.6290	0.6369	0.6363	0.6595*	**0.6702***
P@20	0.3812	0.4259	0.7816	0.7967	0.7900	0.8082*	**0.8093***

The cosine similarity approaches presented by Xu et al. [13], $\cos_{tf\text{-}idf}$ and $\cos_{\mathrm{bm25}}$, have much lower performance values than the rest of the approaches, even though Xu et al. report for them a performance better than the tf approach, presented by Noll and Meinel [6]. A possible reason for this contradiction is the difference between Xu et al.'s and our evaluation setups. On one hand, the length normalization factor used in the cosine similarity function penalizes those documents with a high amount of assigned tags, i.e., those documents that are more popular, in favor of documents that have fewer related tags. This penalization factor may be self-defeating according to different studies [2, 4] which suggest that a popular document has a higher chance to be relevant to a user. We noticed that the documents returned by the Web search system were highly diverse in terms of popularity, and thus the discrimi-

[11] http://developer.yahoo.com/search/boss/

nation of popular documents had a sensible negative impact. On the other hand, Xu et al. make use of a controlled document collection, no larger than 15K documents, which may not have these characteristics.

The $bm25_{u_m}$ approach obtains a performance statistically significant higher than the $bm25_{d_n}$ approach. This implies that, in a folksonomy model, it is better to assume that the user acts as a document in terms of the probabilistic model's relevance computation. Compared to the other personalization approaches, the $bm25_{u_m}$ approach has a better performance in terms of the MRR metric, outperforming the tf approach, which is the best found in the state of the art. However, it has a performance statistically significant lower than the $tf\text{-}if$ approach. The performance of $tf\text{-}if$ approach is higher than both the if and the $bm25$ approaches, with statistical significant differences on all the used metrics. These results highlight the importance of incorporating the global frequencies calculated for a given tag, i.e., the tag user inverse frequency iuf and the tag document inverse frequency idf.

Moreover, since the $bm25_{u_m}$ and the $tf\text{-}if$ approaches are based on different models, the probabilistic and the vector space models, respectively, we investigate the performance of a combination of both approaches. We use a simple, parameter free aggregation strategy, CombSUM with rank-based normalization [10], to merge their rankings. The obtained performance results are presented in the last column of Table 2, and are encouraging: this strategy is the highest performing approach, achieving a 13.91% improvement on MRR, and a 8.12% improvement in terms of P@5, with respect to the best performing state of the art approach, tf, indicating that both our approaches complement each other. We also computed Kendall's tau over the ranks produced by our combination and tf approaches, in order to check if these techniques were personalizing results differently. The average Kendall's tau over all topics was 0.185 (SD = 0.187, $p < 0.05$) which lead us to think that this was the case. Other combinations did not result in a performance improvement.

5.2 Results of Folksonomy-based Personalized Web Search

We now investigate the performance of the personalization approaches when used in combination with a Web search system. In order to do this, we merge the result lists returned by the Web search system (denoted as $S(q)$ in Section 3) with the result lists produced by the personalization approaches, i.e., the results evaluated in the previous section.

Table 3. Personalization approaches performance when combined with the Web search engine results. Values with an asterisk indicate a statistically significant higher value than the Web search ranking (Wilcoxon test, $p < 0.05$). Values marked with a † also indicate a statistically significant higher value than the tf approach. Values in bold are the highest significant values

Metric	baseline	$\cos_{tf\text{-}idf}$	$\cos_{\mathrm{bm25}}$	$bm25_{d_n}$	tf	$bm25_{u_m}$	$tf\text{-}if$	comb
MRR	0.3292	0.1626	0.1810	0.3750*	0.3905*	0.3931*	0.4019†	**0.4073**†
P@5	0.4523	0.2354	0.2696	0.5435*	0.5554*	0.5593*	0.5652†	**0.5705**†
P@10	0.5793	0.3968	0.4325	0.6720*	0.6859*	0.6790*	0.6903†	**0.6955**†
P@20	0.7078	0.5945	0.6181	0.7903*	0.7952*	0.7983*	0.7980*	**0.8006**†

As a baseline, we use the Web search system. In order to make a fairer comparison, we eliminate from its result lists those documents that were not bookmarked by any user. The final ranked lists are combinations of both the non-personalized and the personalized rank lists using CombSUM with rank based normalization as aggregation method [10]. Table 3 shows the performance values of the personalization approaches combined with the Web search. Values are correlated with those presented in Table 2.

The cosine similarity personalization approaches degrade the performance of the Web search, while the rest of approaches outperform the baseline. The two approaches proposed in this work, $tf\text{-}if$ and $bm25_{u_m}$, perform better than both the baseline and tf. Again, the combination of tf and $bm25_{u_m}$ personalization approaches yield the best performance, both in terms of MRR and precision. This demonstrates the complementarily of both approaches, whose combination achieves 23.72% and 4.3% improvements with respect to the baseline and tf approaches, respectively.

6 Conclusions and Future Work

In this paper, we have presented two novel techniques that exploit user and document profiles defined in a social tagging system to personalize the result rankings of a Web search system. The first personalization approach is based on the vector space information retrieval model, and incorporates the concepts of tag inverse document frequency and tag inverse user frequency, which are global measures that rely on the tag distribution within the folksonomy-based user and document profiles. The second personalization approach is an adaptation of the BM25 probabilistic model to folksonomy systems based on the above user and document representations.

We have also proposed a novel evaluation framework and a topic generation methodology which allow the automatic evaluation of folksonomy-based Web search personalization approaches. The results obtained with the evaluations conducted over a dataset from Delicious social bookmarking system show that our techniques outperform the state of the art folksonomy-based personalization approaches. Furthermore, we demonstrate how our two personalization techniques can complement each other, achieving the best overall performance when combined by a well-known rang aggregation strategy. We claim that the key points of the achieved performances are 1) the use of the proposed global tag importance measures, 2) the removal of length normalization factors in personalization formulas, and 3) the adaptation of the probabilistic model.

The presented techniques can be applied to any Web search system, providing personalization capabilities to any user who has a profile in a social tagging service. Thus, with no extra effort, a user can personalize and enhance the results provided by a certain Web search engine. In our evaluations, we obtained a performance increase of 23.7% over Yahoo! Search, demonstrating the feasibility of this personalization paradigm.

The approaches evaluated in this paper exploit the folksonomy's user, tag and document distribution. However, there are also specific techniques which exploit the folksonomy structure in order to expand the folksonomy-based profiles. The main problem is that, to date, these techniques are not easily scalable to the Web, and have

to be evaluated in small controlled collections [13]. Thus, we were unable to incorporate them into our Web search personalization framework. In the future, we will investigate a scalable expansion strategy that could allow its application to personalization approaches focused on Web search.

We will also study how our personalization techniques can be combined with folksonomy-based popularity measures presented in the state of the art [2, 4]. Although our techniques include some basic document popularity factors, they could be complemented by the above more complex measures.

Acknowledgements. This research was supported by the European Commission under contract FP6-027122-SALERO and by the Spanish Ministry of Science and Education (TIN2008-06566-C04-02).

References

1. Au-Yeung, C. M., Gibbins, N., Shadbolt, N.: A study of user profile generation from folksonomies. In: Proc. of the Social Web and Knowledge Management Workshop (2008)
2. Bao, S., Xue, G., Wu, X., Yu, Y., Fei, B., Su, Z.: Optimizing web search using social annotations. In: Proc. of WWW 2007, pp. 501-510. ACM Press, New York (2007)
3. Chirita, P. A., Costache, S., Nejdl, W., Handschuh, S.: P-tag: large scale automatic generation of personalized annotation tags for the web. In: Proc. of WWW 2007, pp. 845-854. ACM, New York (2007)
4. Hotho, A., Jäschke, R., Schmitz, C., Stumme, G.: Information retrieval in folksonomies: search and ranking. The Semantic Web: Research and Applications. LNCS, vol. 4011, pp. 411-426. Springer, Heidelberg (2006)
5. Micarelli, A., Gasparetti, F., Sciarrone, F., Gauch, S.: Personalized search on the World Wide Web. The Adaptive Web. LNCS, vol. 4321, pp. 195-230. Springer, Heidelberg (2007)
6. Noll, M. G., Meinel, C.: Web search personalization via social bookmarking and tagging. In: Proc. of ISWC 2007. LNCS, vol. 4825, pp. 367-380. Springer, Heidelberg (2007)
7. Jansen, B. J., Spink, A., Bateman, J., Saracevic, T.: Real life information retrieval: a study of user queries on the web. SIGIR Forum 32 (1), 5-17 (1998)
8. Robertson, S. E. Walker, S: Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval. In: Proc. of SIGIR 2004, pp. 345-354. Springer (2004)
9. Shepitsen, A., Gemmell, J., Mobasher, B., and Burke, R.: Personalized recommendation in social tagging systems using hierarchical clustering. In: Proc. of RecSys 2008, pp. 259-266. ACM Press, New York (2008)
10. Shaw, J. A. and Fox, E. A. Combination of multiple searches. In: Text REtrieval Conference, pp. 243-252 (1993)
11. Szomszor, M., Alani, H., Cantador, I., O'hara, K., Shadbolt, N.: Semantic modelling of user interests based on cross-folksonomy analysis. In: Proc. of ISWC 2008. LNCS, vol. 5318, pp. pp. 632-648. Springer, Heidelberg (2008)
12. Voorhees, E.: The TREC-8 question answering track report. In: The 8th Text REtrieval Conference (TREC 8), pp. 77-82 (1999)
13. Xu, S., Bao, S., Fei, B., Su, Z., Yu, Y.: Exploring folksonomy for personalized search. In: Proc. of SIGIR 2008, pp. 155-162, ACM Press, New York (2008)