Exploring Folksonomy for Personalized Search

Shengliang Xu*
Shanghai Jiao Tong University
Shanghai, 200240, China
slxu@apex.sjtu.edu.cn

Shenghua Bao*
Shanghai Jiao Tong University
Shanghai, 200240, China
shhbao@apex.sjtu.edu.cn

Ben Fei
IBM China Research Lab
Beijing, 100094, China
feiben@cn.ibm.com

Zhong Su
IBM China Research Lab
Beijing, 100094, China
suzhong@cn.ibm.com

Yong Yu
Shanghai Jiao Tong University
Shanghai, 200240, China
yyu@apex.sjtu.edu.cn

ABSTRACT

As a social service in Web 2.0, folksonomy provides the users the ability to save and organize their bookmarks online with "social annotations" or "tags". Social annotations are high quality descriptors of the web pages' topics as well as good indicators of web users' interests. We propose a personalized search framework to utilize folksonomy for personalized search. Specifically, three properties of folksonomy, namely the *categorization*, *keyword*, and *structure* property, are explored. In the framework, the rank of a web page is decided not only by the term matching between the query and the web page's content but also by the topic matching between the user's interests and the web page's topics. In the evaluation, we propose an automatic evaluation framework based on folksonomy data, which is able to help lighten the common high cost in personalized search evaluations. A series of experiments are conducted using two heterogeneous data sets, one crawled from Del.icio.us and the other from Dogear. Extensive experimental results show that our personalized search approach can significantly improve the search quality.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Information search and Retrieval—*Search Process*

General Terms

Algorithms, Measurement, Experimentation, Performance

Keywords

Folksonomy, Personalized Search, Topic Space, Web 2.0, Automatic Evaluation Framework

*Part of this work was done while Shengliang Xu and Shenghua Bao were interns at IBM China Research Lab.

SIGIR'08, July 20–24, 2008, Singapore.
Copyright 2008 ACM 978-1-60558-164-4/08/07 ...$5.00.

1. INTRODUCTION

In today's search market, the most popular search paradigm is keyword search. Despite simplicity and efficiency, keyword queries can not accurately describe what the users really want. People engaged in different areas may have different understandings of the same literal keywords. Authors of [26], concluded that people differ significantly in the search results they considered to be relevant for the same query.

One solution to this problem is *Personalized Search*. By considering user-specific information [21], search engines can to some extent distinguish the exact meaning the users want to express by the short queries. Along with the evolution of the World Wide Web, many kinds of personal data have been studied for personalized search, including user manually selected interests [16, 8], web browser bookmarks [23], users' personal document corpus [7], search engine click-through history [10, 22, 24], etc. In all, search personalization is one of the most promising directions for the traditional search paradigm to go further.

In recent years, there raises a growing concern in the new Web 2.0 environment. One feature of Web 2.0 that distinguishes it from the classical World Wide Web is the social data generation mode. The service providers only provide platforms for the users to collaborate and share their data online. Such services include folksonomy, blog, wiki and so on. Since the data are generated and owned by the users, they form a new set of personal data. In this paper, we focus on exploring folksonomy for personalized search.

The term "folksonomy" is a combination of "Folk" and "Taxonomy" to describe the social classification phenomenon [3]. Online folksonomy services, such as Del.icio.us , Flickr and Dogear [19] , enable users to save and organize their bookmarks, including any accessible resources, online with freely chosen short text descriptors, i.e. "social annotations" or "tags", in flat structure. The users are able to collaborate during bookmarking and tagging explicitly or implicitly. The low barrier and facility of this service have successfully attracted a large number of users to participate.

The folksonomy creates a social association between the users and the web pages through social annotations. More specifically, a user who has a given annotation may be interested in the web pages that have the same annotation. Inspired by this, we propose to model the associations between the users and the web pages using a topic space. The interests of each user and the topics of each web page can be mapped to vectors in the topic space. The personalized search is conducted by ranking the web pages in two guidelines, term matching and topic matching. When a user u issues a query q, a web page p is ranked not only by the term similarity between q and p but also by the topic similarity between u and p. The social annotations in folksonomy naturally form a social topic space. Three properties of folksonomy are studied for the topic space estimation:

The *categorization* property. Many of the social annotations are subject descriptor keywords at various levels of specificity [17]. The selection of proper annotations for a web page is somewhat a classification of the web page to the categories represented by the annotations.

The *keyword* property. As discussed in [12, 4, 27], the annotations can be seen as good keywords for describing the respective web pages from various aspects.

The *structure* property. In folksonomy systems, users' bookmarking actions form a cross link structure between the users and the web pages. Since all the folksonomy data are publicly available, the structure can be fully explored.

Some of the prior studies show similar ideas. In [8, 13, 22, 21, 16], they use ODP taxonomy structure to represent the topics of the web pages and the interests of the users. As a comparison, we also apply ODP in our work to show whether or not the classical web page taxonomy still perform well enough for the Web 2.0 search personalization.

As for evaluation, we propose a new evaluation framework for personalized search using folksonomy data. The framework is low cost. Thus it is able to help lighten the common high barrier in personalized search evaluation. Extensive experimental results show that our personalized search algorithm outperforms the baselines significantly.

The rest of this paper is organized as follows. Section 2 lists some related work. In Section 3, after a detailed analysis of folksonomy, the personalized search algorithms are discussed. Section 4 presents the novel personalized search evaluation framework. In Section 5, we report the experiment results. Section 6 lists some discussions about our work. Finally, we conclude our work and list some future work in Section 7.

2. RELATED WORK

This paper brings together two areas, personalized search and folksonomy, both of which already exist a lot of prior efforts. In this section we present a separate review on either of them.

2.1 Personalized Search

As early as in 2000, Lawrence [15] pointed out that next-generation search engines will increasingly use context information to improve search effectiveness. In 2002, Pitkow et al. further identified two primary strategies, query refinement and result processing, to personalize search in [21].

Query Refinement, also called *Query Expansion*, refers to the modification to the original query, including augmenting the query by other terms or changing the original weight of each query term. Much work has been done in this area, like [25, 7], etc. However, since our work focuses on result processing, these prior efforts are not relevant to us closely. We do not review them in detail here.

Result Processing includes *result reranking* according to each user's personal needs, *result clustering* for better presentation, etc. Among these, result reranking is one of the most widely used. Haveliwala in [13] proposed to calculate a set of PageRanks for each web page biased on the top most 16 ODP categories. The ODP categories in his work is a little similar to the topic space we will propose in the personalized search framework. But our topic space is much more general than their ODP categories. Further in [22], Qiu and Cho proposed a sophisticated approach to build user models from user click history and combine it with Haveliwala's work for personalized search. In some other studies, such as [16, 8, 21] the ODP category structure is also accepted for modeling the web pages' topics and the users' interests. The ODP categories in these studies is a little similar to the topic space we will propose in the personalized search framework but our topic space is more general. In a recent study [20], Noll and Meinel proposed to rerank the non-personalized search results by considering the user's social annotations and the search results' social annotations. Their work is rather simple while effective. The success they achieved is a strong support for our work. Recently, Dou et al. [10] proposed an evaluation framework for personalized search using user click-through history, which needs a lot of user click through data from a real life search engine. Though the technology sounds promising, it is unpractical for most of the researchers because the click through data of search engines are not publicly accessible.

Except the above, there are still a lot of wonderful prior studies on result refinement for personalized search, such as [24, 25], etc. Since they are not very relevant to our work, we don't present the detailed reviews here.

2.2 Folksonomy

Existing research on folksonomy can be mainly divided into two directions. The first is the survey and analysis of the general characteristics of folksonomy systems. The second is the exploring of folksonomy for various applications.

The semantic values of folksonomy. In [17] , the authors investigated two of the most famous folksonomy service providers Del.icio.us and Flickr and gave the strengths and weaknesses of annotation data. Golder & Huberman gave a deep investigation of the Del.icio.us tag data in [12]. Al-Khalifa & Davis analyzed the semantic value of social annotations and got the conclusion that the folksonomy tags are semantically richer than keywords extracted using a major search engine extraction service like Yahoo TE [2].

The collaborative link structure. Several prior efforts propose to model the underlying link structure of folksonomy by graphs. In [14], the authors viewed the tagging system as a tripartite network with users, tags and URLs as three kinds of nodes. Catutto et al. investigated the underlying tripartite graph of the tagging systems in [5]. They concluded that folksonomies exhibit a small world structure.

Applications. Many applications of social annotations have been carried out in recent years, most of which focus on exploring the semantic value of annotations. [27] and [18] both exploited the latent semantics under the tag literature. Bao et al. in [4] proposed to measure the similarity and popularity of web pages from web users' perspective by calculating SocialSimRank and SocialPageRank, respectively.

3. USING FOLKSONOMY FOR PERSONALIZED SEARCH

In this section, we first give a short analysis of folksonomy, and then discuss in detail the approach we propose for personalized search.

3.1 Analysis of Folksonomy

What folksonomy can bring us in personalized search? The best way to answer this question is to analyze it.

Social Annotations as Category Names. In the folksonomy systems, the users are free to choose any social annotations to classify and organize their bookmarks. Though there may be some noise, each social annotation represents a topic that is related to its semantic meaning [17]. Based on this, the social annotations owned by the web pages and the users reflect their topics and interests respectively.

Social Annotations as Keywords. As discussed in [2, 4, 27] the annotations are very close to human generated keywords. Thus, the social annotations usually can well describe or even complement the content of the web pages.

Collaborative Link Structure. One of the most important benefits that online folksonomy systems bring is the collaborative link structure created by the users unconsciously. The underlying link structure of the tagging systems has been explored in many prior efforts [4, 18, 27]. The whole underlying structures of folksonomy systems are rather complex. Different researchers may reduce the complexity of modeling the structure by various simplified model, e.g. in [27], the structure is modeled through a latent semantic layer while in [4] the relations between the annotations and the web pages are modeled using a bipartite graph. In our work, since the relations between the users and the web pages are very important, we model the structure using a user-web page bipartite graph as shown in Figure 1.

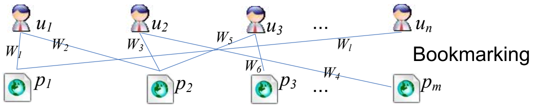

Figure 1: User-web page bipartite structure

where u_i, $i = 1, 2, \cdots, n$ denote n users, p_j, $j = 1, 2, \cdots, m$ denote m web pages, W_k, $k = 1, 2, \cdots, l$ are the weights of the links, i.e. the bookmarking actions of the users. One of the simplest implementation of the weights is the number of annotations a user assigned to a web page.

3.2 A Personalized Search Framework

In the classical non-personalized search engines, the relevance between a query and a document is assumed to be only decided by the similarity of term matching. However, as pointed in [21], relevance is actually relative for each user. Thus, only query term matching is not enough to generate satisfactory search results for various users.

In the widely used Vector Space Model(VSM), all the queries and the documents are mapped to be vectors in a universal term space. The similarity between a query and a document is calculated through the cosine similarity between the query term vector and the document term vector. Though simple, the model shows amazing effectiveness and efficiency.

Inspired by the VSM model, we propose to model the associations between the users and the web pages using a *topic space*. Each dimension of the topic space represents a topic. The topics of the web pages and the interests of the users are represented as vectors in this space. Further we define a topic similarity measurement using the cosine function. Let $\vec{p_{ti}} = (w_{1,i}, w_{2,i}, \cdots, w_{\alpha,i})$ be the topic vector of the web page p_i where α is the dimension of the topic space and $w_{k,i}$ is the weight of the k^{th} dimension. Similarly, let $\vec{u_{tj}} = (w_{1,j}, w_{2,j}, \cdots, w_{\alpha,j})$ be the interest vector of the user u_j. The topic similarity between p_i and u_j is calculated as Equation 1.

$$sim_{topic}(p_i, u_j) = \frac{\vec{p_{ti}} \bullet \vec{u_{tj}}}{|\vec{p_{ti}}| \times |\vec{u_{tj}}|} \quad (1)$$

Based on the topic space, we make a fundamental personalized search assumption, i.e. Assumption 1.

Assumption 1. *The rank of a web page p in the result list when a user u issues a query q is decided by two aspects, a term matching between q and p and a topic matching between u and p.*

When a user u issues a query q, we assume two search processes, a term matching process and a topic matching process. The term matching process calculates the similarity between q and each web page to generate a user unrelated ranked document list. The topic matching process calculates the topic similarity between u and each web page to generate a user related ranked document list. Then a merge operation is conducted to generate a final ranked document list based on the two sub ranked document lists. We adopt ranking aggregation to implement the merge operation.

Ranking Aggregation is to compute a "consensus" ranking of several sub rankings [11]. There are a lot of rank aggregation algorithms that can be applied in our work. Here we choose one of the simplest, Weighted Borda-Fuse (WBF). Equation 2 shows our idea.

$$r(u, q, p) = \gamma \cdot r_{term}(q, p) + (1 - \gamma) \cdot r_{topic}(u, p) \quad (2)$$

where $r_{term}(q, p)$ is the rank of the web page p in the ranked document list generated by query term matching, $r_{topic}(u, p)$ is the rank of p in the ranked document list generated by topic matching and γ is the weight that satisfies $0 \leq \gamma \leq 1$.

Obviously, how to select a proper topic space and how to accurately estimate the user interest vectors and the web page topic vectors are two key points in this framework. The next two subsections discuss these problems.

3.3 Topic Space Selection

In web page classification, the web pages are classified to several predefined categories. Intuitively, the categories of web page classification are very similar to the topics of the topic space. In today's World Wide Web, there are two classification systems, the traditional taxonomy such as ODP and the new folksonomy. The two classification systems can be both applied in our framework. Since our work focuses on exploring the folksonomy for personalized search, we set the ODP topic space as a baseline.

3.3.1 Folksonomy: Social Annotations as Topics

Based on the *categorization* feature, we set the social annotations to be the dimensions of the topic space. Thus, the topic vector of a web page can be simply estimated by its social annotations directly. In the same way, the interest vector of a user can be also simply estimated by her social annotations.

Obviously, if we treat the users and the web pages as documents, the social annotations as terms, the above setting is right the VSM. Since the VSM has developed for a long time, there have been a large number of mature technologies to improve the VSM search effectiveness. All these can be easily applied here. One of the most important in VSM is the weighting for document terms. Similarly, the topic weighting here is also very important. The simplest while widely used one is $tfidf$.

$$w = tf \times \log \frac{N}{n_i}, \tag{3}$$

where tf denotes the term frequency, N denotes the total number of documents in the whole collection and n_i denotes the number of documents in which the term appears. Beside this, BM25 weighting scheme is a more sophisticated alternative, which represents state-of-the-art retrieval functions used in document retrieval

$$w = \log \frac{N - n_i + 0.5}{n_i + 0.5} \cdot \frac{tf \cdot (k_1 + 1)}{tf + k_1 \cdot (1 - b + b \cdot \frac{dl}{avgdl})}, \tag{4}$$

where k_1 and b are free parameters, dl denotes the document length and $avgdl$ denotes the average document length of all the documents in the collection.

3.3.2 Taxonomy: ODP Categories as Topics

In web page taxonomy, the “DMOZ” Open Directory Project (ODP) is the largest, most comprehensive human-edited directory of the web. This high quality and free web taxonomy resource has been used in rather a number of prior researches like [21, 8, 13, 22, 16]. Some of these studies show similar idea as ours, especially [13] and [22]. They use the ODP categories as topics to calculate a set of topic biased PageRanks, which are used in personalized search. Following their steps, we can also choose ODP’s 16 top categories as the dimensions of the topic space. However, 16 categories may be too few for our personalized search task comparing to the folksonomy categories. Thus, we make another choice of totally 1171 categories, including all the second level categories of ODP and the third level categories of TOP/Computers. The choice is based on the consideration that the data corpus we will use in experiments are mostly about computer science.

Now the question is how to estimate the topic vectors and interest vectors. ODP releases all the data in RDF format. In the RDF file, each of the web pages included in ODP attaches a short description. All the descriptions of the web pages under a category can be merged to create a term vector of the corresponding category. Then the topic vector of a web page can be calculated by cosine similarity of the category’s term vector and the social annotations of the web page. Similarly, the interest vector of a user can be calculated by cosine similarity of the category’s term vector and the social annotations owned by the user.

3.4 Interest and Topic Adjusting via Bipartite Collaborative Link Structure

In Section 3.1, we have modeled the underlying collaborative structure of a folksonomy system as a bipartite graph. The bipartite structure is the result of user collaboration which is one of the main advantages that online folksonomy service over offline desktop bookmarks. Intuitively, the topics of the web pages that a user saved in social tagging systems exhibit the user’s interests. In return, the interests of the users who saved a given web page also imply the topics of the web page to some extent. Furthermore, it’s not difficult to infer that this process is actually iterative. We propose to fully explore this bipartite structure for adjusting the initial estimation of users’ interest vectors and the web pages’ topic vectors using an iterative algorithm.

Formally, Let $G = (V, E)$ be the graph, where the nodes in V represent users and web pages, and the edges E represent the bookmarking actions. The nodes in V are divided into two subsets $U = \{u_1, u_2, \cdots, u_n\}$ representing the users and $P = \{p_1, p_2, \cdots, p_m\}$ representing the web pages. In Table 1, we list all the symbols we will use in the algorithm.

Table 1: Symbols used in the Topic Adjusting Algorithm

Symbol	Meaning
W	The adjacency matrix, in which the rows represent the users and the columns represent the web pages. $W_{i,j}$ is set to the number of annotations that u_i gives to p_j.
W_{rn}	The row normalized version of W.
W_{cn}	The column normalized version of W.
$r_{i,j}$	The j^{th} normalized interest of the i^{th} user
$t_{i,j}$	The j^{th} normalized topic of the i^{th} web page
R	The row normalized interest matrix of all the users, in which the rows represent the users and the columns represent the interests. $R_{i,j}$ is the j^{th} interest value of u_i
T	The row normalized topic matrix of all the web pages, in which the rows represent the web pages and the columns represent the topics. $T_{i,j}$ is the j^{th} topic value of p_i
α	The weight of the initial estimated user interest
β	The weight of the initial estimated web page topic

Each iteration of this algorithm is performed in two steps.

1) User interest adjusting by related web pages.

$$r_{i,j} = \alpha \cdot r^0_{i,j} + (1 - \alpha) \cdot \frac{\sum_{k=1}^{m} t_{k,j} \cdot W_{i,k}}{\sum_{k=1}^{m} W_{i,k}} \tag{5}$$

where $r^0_{i,j}$ is the initial value of $r_{i,j}$.

2) Web page topic adjusting by related users.

$$t_{i,j} = \beta \cdot t^0_{i,j} + (1 - \beta) \cdot \frac{\sum_{k=1}^{n} r_{k,j} \cdot W_{k,i}}{\sum_{k=1}^{n} W_{k,i}} \tag{6}$$

where $t^0_{i,j}$ is the initial value of $t_{i,j}$.

As we can see from the above two equations, we reserve in each iteration an α and a β weight of the initial interest value and the initial topic value respectively. The reason is that, since $r^0_{i,j}$ and $t^0_{i,j}$ are estimated directly from the social annotations’ literal contents while $(\sum_{k=1}^{m} t_{k,j} \cdot W_{i,k})/\sum_{k=1}^{m} W_{i,k}$ and $(\sum_{k=1}^{n} r_{k,j} \cdot W_{k,i})/\sum_{k=1}^{n} W_{k,i}$ are from the link structure, they are two heterogeneous parts. The two weights, α and β, are to reserve the influence of the social annotations’ literal contents in the final adjusted vectors.

Besides, though the forms of the above two equations seem to be complicated, the operations are actually linear combination. Thus the topic vectors of the web pages and the interest vectors of the users must be in the same scale. Thus, before the running of the algorithm we normalize all the vectors.

Finally, the above two equations can be rewritten in the form of matrices as following:

$$R^{t+1} = \alpha R^0 + (1 - \alpha) W_{rn} T^t \tag{7}$$

$$T^{t+1} = \beta T^0 + (1 - \beta) W_{cn}^T R^{t+1} \tag{8}$$

We claim that this iterative algorithm converges to a fixed point finally. In the following we give a short proof. We don't list the detailed analysis of this algorithm because of page limitation. The interested readers can refer to some prior studies such as [30] [29], inspired from which we have the idea of this algorithm.

Proof. Without loss of generality, we only prove R^i can converge to a fixed point. Let W_α be $(1-\alpha)W_{rn}$ and W_β be $(1-\beta)W_{cn}^T$, we can expand Equation 7 as following:

$$\begin{aligned} R^{i+1} &= \alpha\{E + W_\alpha W_\beta + (W_\alpha W_\beta)^2 + \cdots + (W_\alpha W_\beta)^{i+1}\}R^0 \\ &+ \beta W_\alpha\{E + W_\beta W_\alpha + (W_\beta W_\alpha)^2 + ...(W_\beta W_\alpha)^i\}T^0 \end{aligned}$$

Thus,

$$\begin{aligned} & \lim_{i\to+\infty} \|R^{i+1} - R^i\| \\ =\; & \lim_{i\to+\infty} \|\alpha[W_\alpha W_\beta]^{i+1}R^0 + \beta W_\alpha(W_\beta W_\alpha)^i T^0\| \\ =\; & \lim_{i\to+\infty} \|\alpha(1-\alpha)^{i+1}(1-\beta)^{i+1}(W_{rn}W_{cn}^T)^{i+1}R^0 \\ +\; & \beta(1-\alpha)^{i+1}(1-\beta)^i(W_{cn}^T W_{rn})^i T^0\| \end{aligned}$$

On the one hand, consider that W_{rn} and W_{cn}^T are both row normalized, they are actually two Markov matrices, thus $W_{rn}W_{cn}^T$, $W_{cn}^T W_{rn}$, $(W_{rn}W_{cn}^T)^{i+1}$ and $(W_{cn}^T W_{rn})^i$ are also Markov matrices. On the other hand, because that $0 < \alpha < 1$ and $0 < \beta < 1$, we can derive:

$$\begin{aligned} \lim_{i\to+\infty}\{\alpha(1-\alpha)^{i+1}(1-\beta)^{i+1}\} &= 0 \\ \lim_{i\to+\infty}\{\beta(1-\alpha)^{i+1}(1-\beta)^i\} &= 0 \end{aligned}$$

Thus, we can finally derive that $\lim_{i\to+\infty} \|R^{i+1} - R^i\| = 0$, i.e. R^i is convergent. □

For convenience, we refer to this algorithm as Topic Adjusting Algorithm in the rest of the paper.

4. AN EVALUATION FRAMEWORK FOR PERSONALIZED SEARCH USING SOCIAL ANNOTATIONS

In the community of personalized search, evaluation is not an easy task. Generally speaking, the evaluation methods used in prior personalized search studies fall into two categories, user experience study [21, 25, 22, 7, 8] and search engine query logs [10, 24].

The user study approach, though widely accepted in most of the prior efforts, needs many users to involve in the experiments, which is a rather high cost. In addition, since the users who take part in the experiments know that they are being tested, they may bias the experiment results. The search engine query logs approach needs a large portion of real life search logs. This is not possible for most of the researchers, including us. The search engine service providers are not willing to release their query logs because they include privacy of the users. In addition, the relevance assumption based on user clicks is strongly biased by the search engines.

Under this condition, we propose a new evaluation framework for personalized search based on social annotations. The main obstacle that raises the difficulty of evaluation for personalized search is that we must have enough user-specific relevance judgement data. In the user experience study, these data are collected from the experiment participants directly. In the search log approach, the researchers make an assumption that the user clicks reflect their relevance judgement. Thus they can collect a lot of experiment data without any extra user efforts. As for our evaluation framework, we make an assumption similar to the search log approach, i.e.

Assumption 2. *The users' bookmarking and tagging actions reflect their personal relevance judgement.*

For example, if a user assigned an annotation "*java*" to the Apache Lucene homepage (http://lucene.apache.org) we assume that the user will consider this web page as relevant if she issues "*java*" as a query. Of course, it's also the truth that a lack of an annotation doesn't necessarily mean irrelevance. However, to the best of our knowledge, this is a common problem for all the prior evaluation approaches for personalized search within the web scale.

This assumption is based on three considerations.

1) In today's search technology, keyword query is the most popular query representation. According to the *keyword* feature of folksonomy, most of the social annotations are keywords of their owner web pages. Thus, the annotations can be considered as queries to some extent.

2) As discussed in Section 3.1, a web page may contain multiple topics. Different users may be interested in different topics of the same web pages. Most likely the users may choose their favorite topics of the web pages to assign some related annotations. In other words, if the social annotations are issued as queries, different users may consider a web page to be relevant to different queries.

3) Different users may choose various terms as social annotations for the same web page. The annotations reflect their personal preference of daily life vocabulary. In other words, the data don't bias for our experiments.

The above three considerations have been analyzed and explored in several prior efforts [2, 1, 4, 12, 17, 18, 19, 20, 27], because of page limitation, we don't list the detailed analysis here. In all, we expect this new evaluation framework to lighten the high barrier of personalized search evaluation.

5. EXPERIMENTS

5.1 Experiment Setup

5.1.1 Data Set

To fully evaluate our personalized search model, we use two heterogeneous data sets. One is crawled from Del.icio.us during May 2006, consisting of 90,300 web pages, 65,080 distinct annotations and 9,813 users. Since this data set is from the web, it reflects the web users' social bookmarking and tagging patterns. The other one is the tagging records of the Dogear tagging system [19] up to July 7th 2007. The data set consists of 179,835 web pages, 47,993 distinct annotations and 5,192 users. This data set reflects the enterprise users' social bookmarking and tagging patterns.

From each data set, we build three test beds according to the number of bookmarks owned by the users, resulting in totally 6 test beds. The 3 test beds built from the Del.icio.us data set are: 1) 100 randomly selected users who own 5 ~ 10 bookmarks and their tagging records, denoted as DEL.5-10; 2) 100 random users who own 80 ~ 100 bookmarks and their tagging records, denoted as DEL.80-100; 3) all the 31 users who own more than 500 bookmarks and their tagging records, denoted as DEL.gt500. The 3 test beds from the Dogear data set are built in the same way as Del.icio.us, denoted as DOG.5-10, DOG.80-100 and DOG.gt500 respectively. The purpose of building the 6 test beds is not only to evaluate the model in the two different environments, i.e. web and enterprise, but also to evaluate it in the situations of different amount of data.

Before the experiments we perform two data preprocessing processes. 1)Several of the annotations are too personal or meaningless, such as "toread", "Imported_IE_Fa-vorites", "system:imported", etc. We remove some of them manually. 2) Some users may concatenate several words to form an annotation , e.g. javaprogramming, java/programming, etc. We split this kind of annotations with the help of a dictionary. Table 2 presents the statistics of the two data sets and the 6 test beds after data preprocessing where "*num.users*" denotes the number of users, "*max.tags*" denotes the maximum number of distinct tags owned by each user, the rest columns have the similar meanings as "*max.tags*". As for each test bed, we randomly split them into 2 parts, a 80% training part and a 20% test part. The training parts are used to estimate the models while the test parts are used for evaluating. All the preprocessed data sets are used in the experiments. No other filtering is conducted.

Table 2: Statistics of the user owned tags and web pages of the experiment data

Data Set	Num. Users	Max. Tags	Min. Tags	Avg. Tags	Max. Pages	Min. Pages	Avg. Pages
Delicious	9813	2055	1	56.04	1790	1	40.35
Dogear	5192	2288	1	47.43	4578	1	46.78
DEL.gt500	31	1133	74	464.42	1790	506	727.55
DEL.80-100	100	456	2	107.51	100	80	88.43
DEL.5-10	100	64	1	18.53	10	5	7.44
DOG.gt500	92	2147	42	543.87	4578	500	999.04
DOG.80-100	85	295	9	126.96	100	80	89.32
DOG.5-10	100	41	2	16.11	10	5	6.99

5.1.2 Personalized Search Framework Implementation

Our personalized search framework needs two separated ranked lists of web pages. In practice, instead of generating two full ranked lists of all the web pages, an alternative approach that costs less is to rerank only the top ranked results fetched by the text matching model. In the experiments, we conduct such reranking based on two state-of-the-art text retrieval model, BM25 and Language Model for IR (LMIR). Firstly, a ranked list by a text retrieval model is generated. Then top 100 web pages in the ranked list are reranked by our personalized search model.

5.1.3 Parameter Setting

Before the experiments, there are three sets of parameters that must be set. The first two parameters are the α and β in the Topic Adjusting Algorithm in Section 3.4. We simply set them both 0.5 to keep the same influence for the initial social annotations' literal contents and the link structure. The second set of parameters are the set of γs in the ranking aggregation when using various search models under various test beds, i.e. Equation 2. We conduct a simple training process to estimate the γs as shown in Procedure 1. The concrete values of γ under each search model and test bed is listed in Table 3. In addition, we set the three parameters k_1, k_3 and b in BM25 1.2, 7 and 0.75 respectively, which are the default parameter scheme in the lemur toolkit[1]. For the LMIR we accept jelinek-mercer smoothing [28] with α set to 0.3.

Procedure 1. Ranking aggregation parameter training process

```
foreach test bed $TB \in$ 6 test beds do
    split the training part of $TB$ into 4 parts $TN_i$, $1 \leq i \leq 4$
    foreach $TN \in TN_i$ do
        training the interest vectors and the topic vectors
        using other 3 training parts
        run the evaluation 11 times using $TN$ with $\gamma$ set to
        0.0, 0.1, $\cdots$, 1.0 respectively
        record the $\gamma$ that leads to the optimal performance
    set the average of the 4 $\gamma$s as the final parameter
```

5.1.4 Baseline Models

In the experiments we select 4 baseline models, one is the non-personalized text matching model using no extra information except for contents, the second is the model using the top 16 ODP categories as topic space which is denoted as "ODP1", the third is the model using 1171 ODP categories as topics which is denoted as "ODP2", and the last is the model proposed in [20], which is actually a simplified case of our personalized search framework when the topic space is set to be folksonomy and the topic matching function is set to simply counting the number of matched annotations. We refer to it as the "AC" model.

5.1.5 Evaluation Metric

The main evaluation metric we used in our work is mean average precision (MAP), which is a widely used evaluation metric in the IR community. More specifically, in our work, we calculate MAP for each user and then calculate the mean of all the MAP values. We refer it as Mean MAP or MMAP.

$$MMAP = \frac{\sum_{i=1}^{N_u} MAP_i}{N_u}$$

where MAP_i represents the MAP value of the i^{th} user and N_u is the number of users.

In addition, we perform t-tests on average precisions over all the queries issued by all the users in each experimental data set to show whether the experimental improvements are statistical significant or not.

5.2 Performance

Table 3 lists all the 120 experimental results. The columns "text", "ODP1", "ODP2", "AC", "f.tfidf" and "f.bm25" denote the non-personalized text model, the 16 top most ODP topic space personalized model, the 1171 ODP topic space personalized model, the AC model, the folksonomy topic space personalized model using tfidf weighting scheme and the folksonomy topic space personalized model using BM25 weighting scheme. The sub columns "B. A." and "A. A." denote "Before Adjusting by link structure" and "After Adjusting by link structure", respectively. The "*"s in the "MMAP" row stand for four significance levels of the t-test, satisfying $0.05 \geq * > 0.01 \geq ** > 0.001 \geq ***$.

As we can see from the table, the 5 personalized search models all outperform the simple text retrieval models sig-

[1] http://www.lemurproject.org/

Table 3: The γ settings, MMAPs and the improvements(imp.) comparing to the non-personalized text retrieval model using various personalized search models under various data sets

\	\	BM25									
	\	text	ODP1		ODP2		AC[20]	f.tfidf		f.bm25	
\	\	\	B. A.	A. A.	B. A	A. A.	\	B. A.	A. A.	B. A.	A. A.
DOG.5-10	γ	\	0.7	0.2	0.0	0.0	\	0.0	0.0	0.0	0.0
	MMAP	0.0268	0.0416**	0.0512**	0.0445*	0.0677***	0.0813***	0.0748***	0.1045***	0.0549***	0.1065***
	imp.	\	55.2%	91.0%	66.0%	152.6%	203.4%	179.1%	289.9%	104.9%	**297.4%**
DOG.80-100	γ	\	0.1	0.3	0.7	0.0	\	0.4	0.0	0.0	0.0
	MMAP	0.0194	0.0210	0.0254*	0.0219	0.0260	0.0427***	0.0366***	0.0492***	0.0327***	0.0670***
	imp.	\	8.2%	30.9%	12.9%	34.0%	120.1%	88.7%	153.6%	68.6%	**245.4%**
DOG.gt500	γ	\	0.9	0.6	0.9	0.0	\	0.1	0.1	0.0	0.0
	MMAP	0.0208	0.0217**	0.0252***	0.0218	0.0229***	0.0270***	0.0293*	0.0352***	0.0285***	0.0395***
	imp.	\	4.3%	21.2%	4.8%	10.1%	29.8%	40.9%	69.2%	37.0%	**89.9%**
DEL.5-10	γ	\	0.7	0.7	0.0	0.0	\	0.0	0.0	0.0	0.0
	MMAP	0.0321	0.0411***	0.0412***	0.0403	0.0480**	0.0718***	0.0606***	0.0793***	0.0470**	0.0855***
	imp.	\	28.0%	28.3%	25.5%	49.5%	123.7%	88.8%	147.0%	46.4%	**166.4%**
DEL.80-100	γ	\	0.8	0.3	0.6	0.6	\	0.2	0.1	0.1	0.0
	MMAP	0.0238	0.0295***	0.0308***	0.0340***	0.0345***	0.0507***	0.0502***	0.0533***	0.0413***	0.0562***
	imp.	\	23.9%	29.4%	42.9%	45.0%	113.0%	110.9%	123.9%	73.5%	**136.1%**
DEL.gt500	γ	\	0.9	0.8	0.7	0.7	\	0.4	0.3	0.2	0.1
	MMAP	0.0355	0.0367**	0.0385***	0.0394**	0.0430***	0.0524***	0.0551***	0.0563***	0.0565***	0.0621***
	imp.	\	3.4%	8.5%	11.0%	21.1%	47.6%	55.2%	58.6%	59.2%	**74.9%**

\	\	LMIR									
	\	text	ODP1		ODP2		AC[20]	f.tfidf		f.bm25	
\	\	\	B. A.	A. A.	B. A	A. A.	\	B. A.	A. A.	B. A.	A. A.
DOG.5-10	γ	\	0.6	0.2	0.6	0.0	\	0.0	0.0	0.0	0.0
	MMAP	0.0277	0.0427*	0.0477**	0.0474**	0.0665***	0.0783***	0.0657***	0.0888***	0.0470**	0.0958***
	imp.	\	54.2%	72.2%	72.2%	140.1%	182.7%	137.2%	220.6%	69.7%	**245.8%**
DOG.80-100	γ	\	0.9	0.5	0.9	0.1	\	0.6	0.0	0.0	0.0
	MMAP	0.0171	0.0190	0.0228**	0.0184	0.0229	0.0385***	0.0295***	0.0427***	0.0283***	0.0600***
	imp.	\	11.1%	33.3%	7.6%	33.3%	125.1%	72.5%	149.7%	65.5%	**250.9%**
DOG.gt500	γ	\	0.9	0.7	0.9	0.8	\	0.5	0.1	0.0	0.0
	MMAP	0.0200	0.0206***	0.0232***	0.0205	0.0212	0.0250***	0.0276***	0.0320***	0.0262***	0.0348***
	imp.	\	3.0%	16.0%	2.5%	6.0%	25.0%	38.0%	60.0%	31.0%	**74.0%**
DEL.5-10	γ	\	0.7	0.6	0.4	0.4	\	0.0	0.0	0.0	0.0
	MMAP	0.0278	0.0361*	0.0367*	0.0356*	0.0410**	0.0657***	0.0536***	0.0720***	0.0441**	0.0799***
	imp.	\	29.9%	32.0%	28.1%	47.5%	136.3%	92.8%	159.0%	58.6%	**187.4%**
DEL.80-100	γ	\	0.7	0.5	0.5	0.5	\	0.1	0.1	0.1	0.0
	MMAP	0.0220	0.0274***	0.0281***	0.0315***	0.0315***	0.0444***	0.0471***	0.0493***	0.0391***	0.0523***
	imp.	\	24.5%	27.7%	43.2%	43.2%	101.8%	114.1%	124.1%	77.7%	**137.7%**
DEL.gt500	γ	\	0.8	0.8	0.8	0.7	\	0.4	0.3	0.2	0.1
	MMAP	0.0298	0.0329***	0.0345***	0.0348***	0.0386***	0.0480***	0.0514***	0.0517***	0.0514***	0.0584***
	imp.	\	10.4%	15.8%	16.8%	29.5%	61.1%	72.5%	73.5%	96.0%	**96.0%**

nificantly. Though the two ODP topic space search models have rather great improvements over the simple text search models, it is not well enough to fully utilize the folksonomy. The ODP2 model outperforms the ODP1 model in nearly all the experiments while the improvements are not so great. In contrast, even the simplest folksonomy topic space model, i.e. the AC model can beat the ODP models with great improvements. A reason for this is that the interests of the users and the topics of the web pages are actually boundless, thus a predefined static topic space such as ODP is not enough. However, the social annotations in folksonomy are dynamic. They can describe the topics and the interests more precisely.

As for the Topic Adjusting Algorithm, comparing the experimental results of the two columns “B. A.” and “A. A.”, it is clear that the algorithm is very effective. All the models with the adjusted vectors beat the corresponding models with non-adjusted vectors.

Besides, among the three folksonomy topic space models, as we have expected, f.bm25 and f.tfidf outperform the AC model significantly. Notice that the adjusted f.bm25 reaches the optimal performance in all the experiments.

As we can see, all the experiments under various amount of data all output promising results. That means our model can handle all the situations of different amount of data. However one strange phenomenon is the search effectiveness seems to reduce when the amount of data increase. We expected the personalized models to increase performance when the amount of data increase. As to this problem, we manually analyzed the tagging data in the two data sets and find a main cause. Generally the social annotations owned by the users who own a small amount of total social annotations are much semantically richer than the social annotations owned by the users who own a relatively large amount of total social annotations. Because most of the users who own many bookmarks, especially those who have more than 500 bookmarks, directly export their desktop bookmarks into the folksonomy systems. The annotations of these bookmarks are not user manually generated and many of them are obviously noise, such as “Imported_IE_Favorites”, “imported_1/14/06”, “system:imported”, “imported”, etc.

6. DISCUSSIONS

Integrating folksonomy systems with search engines. One problem in implementing our personalized search algorithm in real life is how to access the folksonomy data of a user when she is searching. This won’t be a problem if the search engines and the folksonomy systems are owned by the same company or organization. Yahoo! has given us a solution to this problem not long ago. The two most well known Web 2.0 social tagging websites, Del.icio.us and Flickr, have been purchased by Yahoo!. Furthermore, many folksonomy websites provide simple search engines themselves. The personalization can be implemented on these search engines.

Sparseness of social annotations. Since the social annotations require the users to create explicitly, many users may be reluctant to maintain such personal data. Though more and more users are now engaged in folksonomy, it’s still a small portion of all the search engine users. How to expand the benefit of our personalized search algorithm to all the search engine users? [9] and [6] give us two potential solutions. In [9], the authors collected tagging data automatically from user click through histories by treating queries as annotations and all the clicked web pages as bookmarks. [6] proposed to automatically generate personalized annotations based on users’ personal document corpus. Both the above approaches can be incorporated in our personalized framework easily. Thus the sparseness of social bookmarks can be lightened to a certain extent.

Folksonomy topic dimension reduction. Similar to the document terms, the synonymy and polysemy problem also exist in social annotations. Dimension reduction is a technology to tackle this problem, including LSI, PLSI, etc. However, these algorithms are rather time and space consuming. In our future work, we will study how to reduce folksonomy dimension efficiently and evaluate the effectiveness using reduced dimensions.

7. CONCLUSIONS AND FUTURE WORK

How to effectively use folksonomy for personalized search in Web 2.0 environment is quite a new problem. The main contributions of this paper can be summarized as following: 1) The proposal of a personalized search framework, in which the users and the web pages are associated by a topic space. 2) The proposal of using the social annotations to modeling the topic space. Specifically, three properties of folksonomy, namely the *categorization*, the *keyword* and the *structure* property, are studied. 3) The proposal of an automatic evaluation framework for personalized search using folksonomy data. The evaluation framework is able to lighten the common high cost problem in personalized search evaluations. 4) The evaluations of our personalized search approach using a Del.icio.us corpus and a Dogear corpus show that our approach outperforms the baselines significantly.

This is just our first trial of leveraging folksonomy for personalized search. There are several possible future extensions as listed in the following. 1) we set text retrieval models as our baselines. The purpose of this choice is to show the pure ability of folksonomy for personalized search. However, today's web search engines already account for much meta information such as link structure, anchor text, etc. in addition to the similarity of a query to a document when ranking. We'll explore some approaches to incorporate these information into our framework. 2) the personalized search framework uses Weighted Borda-Fuse as the rank aggregation approach. This simple method is essentially a linear combination. We'll try more sophisticated rank aggregation methods to test the personalized search framework. 3) as for the evaluation framework, we'll test it in some other contexts to show its detailed pros and cons.

8. ACKNOWLEDGEMENT

The authors would like to thank IBM China Research Lab for its continuous support to and cooperation with Shanghai Jiao Tong University. We would also like to express our gratitude to D.R. Millen and J. Feinberg from IBM Watson Research Center for providing us the Dogear social tagging data corpus [19]. Besides, we also appreciate the valuable suggestions of Feng Yun, Mianwei Zhou, and Jinwen Guo. In the end, we would like to thank the anonymous reviewers for their elaborate and helpful comments.

9. REFERENCES

[1] H. S. Al-Khalifa and H. C. Davis. Measuring the semantic value of folksonomies. *Innovations in Inf. Tech.*, pages 1–5, 2006.

[2] H. S. Al-Khalifa and H. C. Davis. Exploring the value of folksonomies for creating semantic metadata. *IJSWIS*, 3(1):13–39, 2007.

[3] G. S. Atomiq. Folksonomy: social classification. http://atomiq.org/archives/2004/08/folksonomy_social_classification.html, August 2004.

[4] S. Bao, G. Xue, X. Wu, Y. Yu, B. Fei, and Z. Su. Optimizing web search using social annotations. In *Proc. of WWW '07*, pages 501–510, 2007.

[5] C. Catutto, C. Schmitz, A. Baldassarri, V. D. P. Servedio, V. Loreto, A. Hotho, M. Grahl, and G. Stumme. Network properties of folksonomies. *AI Communications Journal, Special Issue on "Network Analysis in Natural Sciences and Engineering"*, 2007.

[6] P. A. Chirita, S. Costache, W. Nejdl, and S. Handschuh. P-tag: large scale automatic generation of personalized annotation tags for the web. In *Proc. of WWW '07*, pages 845–854, 2007.

[7] P. A. Chirita, C. S. Firan, and W. Nejdl. Personalized query expansion for the web. In *Proc. of SIGIR '07*, pages 7–14, 2007.

[8] P. A. Chirita, W. Nejdl, R. Paiu, and C. Kohlschütter. Using odp metadata to personalize search. In *Proc. of SIGIR '05*, pages 178–185, 2005.

[9] P. A. Dmitriev, N. Eiron, M. Fontoura, and E. Shekita. Using annotations in enterprise search. In *Proc. of WWW '06*, pages 811–817, 2006.

[10] Z. Dou, R. Song, and J.-R. Wen. A large-scale evaluation and analysis of personalized search strategies. In *Proc. of WWW '07*, pages 581–590, 2007.

[11] C. Dwork, R. Kumar, M. Naor, and D. Sivakumar. Rank aggregation methods for the web. In *Proc. of WWW '01*, pages 613–622, 2001.

[12] S. A. Golder and B. A. Huberman. Usage patterns of collaborative tagging systems. *Journal of Information Science*, 32(2):198–208, 2006.

[13] T. H. Haveliwala. Topic-sensitive pagerank. In *Proc. of WWW '02*, pages 517–526, 2002.

[14] R. Lambiotte and M. Ausloos. Collaborative tagging as a tripartite network. *arXiv:cs.DS/0512090 v2 29 Dec 2005.*

[15] S. Lawrence. Context in web search. *IEEE Data Engineering Bulletin*, 23(3):25–32, 2000.

[16] Z. Ma, G. Pant, and O. R. L. Sheng. Interest-based personalized search. *ACM Trans. Inf. Syst.*, 25(1):5, 2007.

[17] A. Mathes. Folksonomies - cooperative classification and communication through shared metadata, http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html,december,2004.

[18] P. Mika. Ontologies are us: A unified model of social networks and semantics. In *Proc. of ISWC '05*, pages 522–536, 2005.

[19] D. R. Millen, J. Feinberg, and B. Kerr. Dogear: Social bookmarking in the enterprise. In *Proc. of CHI '06*, pages 111–120, 2006.

[20] M. G. Noll and C. Meinel. Web search personalization via social bookmarking and tagging. In *Proc. of ISWC '07*, November 2007.

[21] J. Pitkow, H. Schütze, T. Cass, R. Cooley, D. Turnbull, A. Edmonds, E. Adar, and T. Breuel. Personalized search. *Commun. ACM*, 45(9):50–55, 2002.

[22] F. Qiu and J. Cho. Automatic identification of user interest for personalized search. In *Proc.of WWW'06*, pages 727–736, 2006.

[23] J. Rucker and M. J. Polanco. Siteseer: personalized navigation for the web. *Communications of the ACM*, 40(3):73–76, 1997.

[24] J.-T. Sun, H.-J. Zeng, H. Liu, Y. Lu, and Z. Chen. Cubesvd: a novel approach to personalized web search. In *Proc. of WWW '05*, pages 382–390, 2005.

[25] J. Teevan, S. T. Dumais, and E. Horvitz. Personalizing search via automated analysis of interests and activities. In *Proc. of SIGIR '05*, pages 449–456, 2005.

[26] J. Teevan, S. T. Dumais, and E. Horvitz. Characterizing the value of personalizing search. In *Proc. of SIGIR '07*, pages 757–758, 2007.

[27] X. Wu, L. Zhang, and Y. Yu. Exploring social annotations for the semantic web. In *Proc. of WWW '06*, pages 417–426, 2006.

[28] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to information retrieval. *ACM Trans. Inf. Syst.*, 22(2):179–214, 2004.

[29] D. Zhou, O. Bousquet, T. Lal, J. Weston, and B. Schölkopf. Learning with local and global consistency. *Advances in Neural Information Processing Systems*, 16:321–328, 2004.

[30] D. Zhou, J. Weston, A. Gretton, O. Bousquet, and B. Schölkopf. Ranking on data manifolds. *Advances in Neural Information Processing Systems*, 16:169 – 176, 2004.