PAPER

Tag-Group Based User Profiling for Personalized Search in Folksonomies

Qing DU[†], ***Member***, **Yu LIU**[†], **Dongping HUANG**[†], **Haoran XIE**[††], **Yi CAI**[†a)], ***and*** **Huaqing MIN**[†], ***Nonmembers***

SUMMARY With the development of the Internet, there are more and more shared resources on the Web. Personalized search becomes increasingly important as users demand higher retrieval quality. Personalized search needs to take users' personalized profiles and information needs into consideration. Collaborative tagging (also known as folksonomy) systems allow users to annotate resources with their own tags (features) and thus provide a powerful way for organizing, retrieving and sharing different types of social resources. To capture and understand user preferences, a user is typically modeled as a vector of *tag:value* pairs (i.e., a tag-based user profile) in collaborative tagging systems. In such a tag-based user profile, a user's preference degree on a group of tags (i.e., a combination of several tags) mainly depends on the preference degree on every individual tag in the group. However, the preference degree on a combination of tags (a tag-group) cannot simply be obtained from linearly combining the preference on each tag. The combination of a user's two favorite tags may not be favorite for the user. In this article, we examine the limitations of previous tag-based personalized search. To overcome their problems, we model a user profile based on combinations of tags (tag-groups) and then apply it to the personalized search. By comparing it with the state-of-the-art methods, experimental results on a real data set shows the effectiveness of our proposed user profile method.

key words: *personalized search, user profiling, folksonomy, tag-group effect*

1. Introduction

In recent years, the collaborative tagging systems, which allow users to annotate the resources with tags, are widely used in many web sites. The rich semantics from user-generated tags have been utilized in various applications such as bookmark collection (Del.icio.us*), movie recommendation (Movielens**) and image sharing (Flickr***). With the increasing amount of user-generated tags and resources, one of the most important issues for these applications is to assist users to find their desired resources.

One main stream to solve this problem is to profile users and resources based on those user-generated tags. The resources and the tags posted by Web users to these systems are supposed to be highly dependent on their interests, and the tags given by users provide rich information for building more accurate user profiles. For these existing works, most profiling methods model a user into a vector space in which each dimension is a *tag:value* pair. For example, in a movie tagging system, a user's profile $\overrightarrow{U}_i$ is modeled as a vector:

$$\overrightarrow{U}_i = (sciencefiction : 1, love : 0.5, action : 0.8, horror : 0.2)$$

The tags in the user vector $\overrightarrow{U}_i$ are generated by the user *i* and the corresponding value reflects the preference degree of the user on a tag.

However, limitations exist in the current tag-based personalized search methods. In some cases, the preference degree on a combination of tags (a tag-group) cannot be simply obtained by linearly combining the preference on each single tag in current tag-based user profiles. A combination of the user's favorite tags may not itself be a favorite for the user. In fact, it may even be undesirable. Current tag-based user profiling methods only reflect the preference of a user on the individual tag and cannot reflect the preference degree of a user on a tag group. Let us take a look at the following example.

Example 1: Suppose that there are two users Bob and Alice. Bob likes the spicy chicken dish annotated by two tags "chicken" and "spicy". Bob also like a dish of mild seafood annotated by "seafood" and "mild". Alice is crazy about a dish of plain chicken which is mild taste (annotated by "chicken" and "mild") and a dish of spicy seafood (annotated by "spicy" and "seafood"). However, Alice does not like spicy chicken dish. In this case, if we adopt current methods to model a user by a tag-based profile, Bob and Alice will have the same tag-based profile $U = (spicy : 1, chicken : 1, mild : 1, seafood : 1)$.

Actually, Bob's taste is different from Alice's but they do have the same tag-based profile in Example 1, which implies that they will obtain the same personalized search result when they issue the same queries. It's apparently unreasonable, since what Bob likes is spicy chicken and not spicy or chicken only. More importantly, Alice does not like spicy chicken dish but the system will return spicy chicken to her according to her tag-based profile. The problem is that tag-based user profile cannot distinguish the different combinations of the tags.

Since users are allowed to use more than one tag to annotate the resource, we believe that a user's preference degree on a resource depends on all the tags annotating the resource given by the user as a whole, but not any one of them

Manuscript received February 17, 2014.
Manuscript revised June 2, 2014.

[†]The authors are with the School of Software Engineering, South China University of Technology, Guangzhou, China.
[††]The author is with the Department of Computer Science, City University of Hong Kong, Hong Kong, China.
a) E-mail: ycai@scut.edu.cn (Corresponding author)

DOI: 10.1587/transinf.2014EDP7053

*http://delicious.com
**http://www.movielens.org
***http://www.flickr.com

individually. This is what we call "tag group effect" (we discuss it in detail in Sect. 3). To solve the above problem, we propose a user profiling method based on the tag-group instead of a single tag, where a tag-group is a combination of tags co-occurring in a user annotation on a resource and is the atomic element that reflects the user preference. Based on the tag-group approach to user profiling, we explore the personalized search in folksonomies. The contributions of our work are listed below.

- We explore "tag-group effect" in detail and reveal the limitation of the tag-based user profiling methods that haven't taken tag-group effect into consideration.
- To solve the problem caused by neglecting tag-group effect, we profile a user based on tag-group instead of a single tag, which can more accurately reflect user preferences on resources in folksonomies.
- To show the effectiveness of the tag-group based user profile, we apply it to personalized search in the collaborative tagging applications.
- We conduct experiment to compare the performance of the proposed user profiling method with other main stream user profiling methods on the MovieLens data set. The experimental results show the effectiveness of the new proposed tag-group based user profile.

The rest of this paper is organized as follows. Section 2 reviews the related work on user profiling in folksonomies as well as their applications to achieve personalized search. Section 3 explores the tag-group effect and focuses on the profiling of the tag-group based user profile. Section 4 applies the tag-group based user profile to personalized search. The experimental results are shown in Sect. 5. Section 6 draws conclusion for this paper and discusses some future research directions.

2. Background and Related Work

In this section, we review the relevant work of collaborative tagging system and personalized search in the folksonomies.

2.1 Collaborative Tagging System

Existing research on collaborative tagging system can be divided into two types. The first one focuses on investigating and analyzing the features of tags. [1] analyzed the tag usage patterns, user activities and annotated resources in collaborative tagging systems. [2] did a comprehensive survey to systems like Del.icio.us, and Last.fm, to discover useful tags for information access. The second type mainly attempts to explore social annotations and link structures in folksonomy for various applications. [3] proposed two algorithms, named as SocialSimRank and SocialPageRank respectively, to explore the latent semantics of tags to optimize web search. A recent work done by [4] did evaluations of various similarity measures on semantics of social tags. In [5], the authors proposed that tag based profiles can be represented by naive, co-occurrence and adaptive approach. However, there are drawbacks of this profiling method, either due to the lack of user input which makes discovering co-occurred tags difficult, or the difference existing in personally preferred vocabulary which results in polysemous and synonymy, and thus negatively influences the precision of learned profiles. [6] proposed a method which automatically generates personalized tags for Web pages. [7] presented their analysis on personal data in folksonomies and investigated how accurately user profiles can be generated from those data. Through the tag-based profiles, personalized search proposed by [8] in collaborative tagging systems became possible and popular, as it can facilitate users greatly to find interested resources.

2.2 Personalized Search

The current strategies of personalized search fall into two categories [9]:

• Query expansion: One is query expansion such as [10], which refers to modifying the original query either by expanding it with other terms, or assigning different weights to the terms in the query.

• Resource re-ranking: Another category is result processing, primarily re-ranking, which adapts the search results to a particular user's preference. Most re-ranking strategies attempt to construct a user profile from the user historical behaviors, and use the profile to filter out resources unmatched with his/her interests. [11] modeled both user profiles and resources as topic vectors from ODP hierarchy, thus the matching between user interest and content can be measured by their vector distance. A personalized PageRank algorithm was proposed in [12], which was a modification to the global PageRank on Web, and the search results were personalized based on the hyperlink structure. Besides learning user profiles based on their own browsing histories, [13] also explored social information to refine search results with the help of like-minded neighbors. [14] did comparisons between various personalization approaches, like click based, profile based, long-term based and short-term ones, and proposed an evaluation framework for the strategies.

2.3 Personalized Search in Collaborative Tagging Systems

With the recent development of collaborative tagging systems, some works are proposed on personalized search in the collaborative tagging environment. Noll and Meinel [15] proposed a simple and effective approach to explore user's and resource's related tags based on term frequency, and re-rank the non-personalized search results based on these tags. Xu et al. [8] proposed topic-based personalized search in folksonomy, in which the personalized search was conducted by ranking the resources based on not only term similarity but also topic similarity. In their work, instead of using term frequency, term frequency-inverse document frequency (TF-IDF) and BM25 are used to construct user and resource profiles. Vallet et al. [16] used different techniques to measure the user-resource similarities and compared the

effect of these techniques. Cai and Li [17] proposed NTF to model user profile and consider the personalized search as a fuzzy satisfaction problem. Han et al. [18] proposed a multiple-level user profile to model a user. Although there are several works to handle personalized search with tag-based user and item profiles, they have some limitations. In next sections, we examine and discuss these limitations.

3. Tag-Group Based User Profiling

In this section, we introduce tag-group and tag-group effect in detail. We model a user based tag-groups and propose a novel tag-group based user profiling method.

3.1 Tag-Group and Tag-Group Effect

As mentioned above, user may use multiple tags to annotate a resource in folksonomies. A user prefers a resource due to the resource containing all the annotated features (tags) instead of any one of them individually. In other words, a user may like the combination of the features, and we call such a combination of user annotating tags on a resource as a tag group. We formally define it as follows.

Definition 1: A **tag-group** j denoted by $\overrightarrow{g_j}$, is a vector of tags, i.e.,

$$\overrightarrow{g_j} = (t_1, t_2, \ldots, t_n)$$

There is an effect of tag-group on affecting user preference on a resource. In some cases, the preference degree on a tag group cannot be simply obtained by linearly combining the preference on each single tag. The combination of favorite tags of a user may not be favorite but even annoying for the user as we illustrated in Example 1. We can conclude a property of tag-group effect as follows.

Property 1: Let $e_{i,j}$ denote the preference degree of a user i on the tag group j, as illustrated in Example 1, the tag-group effect means a user i's preference degree on a tag-group j is not always in direct proportion to that of the sum of any j's sub tag-groups, i.e.,

$$e_{i,j} \propto \sum_{t_k \in \overrightarrow{g_j}} e_{i,k}$$

where $\overrightarrow{g_j}$ represents a tag group, t_k is a single tag in tag group j and $e_{i,k}$ is the preference degree of user i on t_k.

Apparently, tag-based user profile doesn't satisfy the tag-group effect property.

3.2 Tag-Group Based User Profiling

Taking tag-group effect into consideration, we model a user profile based on a tag-group instead of individual tags. Different from tag-based user profiling methods, the tag-group based user's profile is defined as follows.

Definition 2: A user profile of user i, denoted by $\overrightarrow{\Psi}_i$, is a vector of tag-group:value pairs, i.e.,

$$\overrightarrow{\Psi}_i = (\overrightarrow{g}_{i,1} : e_{i,1}, \overrightarrow{g}_{i,2} : e_{i,2}, \ldots, \overrightarrow{g}_{i,n} : e_{i,n})$$

where $e_{i,x}$ is the preference degree of user i on $\overrightarrow{g}_{i,x}$. To construct the tag-group based user profile, we need to solve two main problems. The first one is how to obtain the tag-groups in $\overrightarrow{\Psi}$, and the second one is how to calculate $e_{i,x}$ for $\overrightarrow{g}_{i,x}$.

To solve the first problem, we utilize all the tags directly given by a user to annotate a specific resource to build up a tag-group in this paper. Since a user probably use more than one tag to annotate a resource, the tags given by the user directly reflect the user's perspective on the resource to some extend. Hence all the tags given by the user to annotate a resource can form a tag-group.

The second problem is how to calculate the preference degree for each tag-group in the user vector. It mainly reflects a user's preference degree on a resource containing these tag-groups. There are some often used methods like TF and BM25. Cai and Li [17] propose a NTF method which is proved to be better than TF and BM25 in most cases, which can be applied to calculate the reference degree for a tag-group, i.e.,

$$e_{i,x} = \frac{N_{i,x}}{N_i} \tag{1}$$

where $N_{i,x}$ is the number of times user i uses tag-group x to annotate resources, and N_i is the number of resources tagged by user i.

Besides, we observe that the improvement speed of a user's preference degree on a tag-group becomes slowly as the times of the tag-group being used by the user increases. In other words, a user i used a tag-group j 10 times and used another tag-group k 20 times, which doesn't mean user i is twice interested in the resource contains k than the one contains j. According to the above observation, we propose an improved method to calculate $e_{i,x}$ by adopting a *log* function to normalize the effect of appearance times.

$$e_{i,x} = \log_{N_i} N_{i,x} \tag{2}$$

We have conducted experiment to compare NTF and Log in Experiments Section 5.

4. Personalized Search

In this section, we focus on applying tag-group based user profile to personalized search. Personalized search is to find out the information that not only satisfy a user's basic information need but also best match his or her personal interests. Generally, a personalized search approach should take both query relevance and user interest relevance into consideration [19].

4.1 Query Relevance Measurement

Query relevance measurement is to find out to what extent resources satisfy a user's basic information need. A user query is usually in the form of a vector of terms.

Definition 3: A **query** issued by user i denoted by $\overrightarrow{q_i}$ is a vector of terms as follows:

$$\overrightarrow{q_i} = (t^q_{i,1}, t^q_{i,2}, \cdots, t^q_{i,m})$$

where $t^q_{i,x}$ is a term, and m is the total number of terms in the query.

For a resource, we define it as similar as in [17].

Definition 4: A **resource profile** of a resource c, denoted by $\overrightarrow{R_c}$ is a vector of tag:value pairs:

$$\overrightarrow{R_c} = (t_{c,1} : w_{c,1}, t_{c,2} : w_{c,2}, \cdots, t_{c,n} : w_{c,n})$$

where $t_{c,x}$ is a tag being used to describe resource c, n is the number of tags used to describe resource c, $w_{c,x}$ is the value to which resource c possesses the tag (feature) $t_{c,x}$, and $w_{c,x}$ can be intuitively obtained as follows:

$$w_{c,x} = \frac{M_{c,x}}{M_c} \tag{3}$$

where $M_{c,x}$ is the number of users using tag x to annotate resource c, and M_c is the total number of users who use tags to annotate resource c. A higher value of $w_{c,x}$ means that tag x is more salient or representative for resource c.

An aggregation function is proposed in [19] for the final personalized relevance score between a resource and a query issued by user i, which can be applied to our case, i.e.,

$$RScore(\overrightarrow{q}_i, \overrightarrow{\Psi}_i, \overrightarrow{R}_j) = \delta \cdot \gamma(\overrightarrow{q}_i, \overrightarrow{R}_j) + (1-\delta) \cdot \theta(\overrightarrow{\Psi}_i, \overrightarrow{R}_j) \tag{4}$$

$\gamma(\overrightarrow{q}_i, \overrightarrow{R}_j)$ is the query relevance function and $\theta(\overrightarrow{\Psi}_i, \overrightarrow{R}_j)$ is the user interest relevance function. $\gamma(\overrightarrow{q}_i, \overrightarrow{R}_j)$ can be measured by the following equation, i.e.,

$$\gamma(\overrightarrow{q}_i, \overrightarrow{R}_j) = \frac{\sum w_{j,x}}{m} \cdot \left(\frac{k}{m}\right)^{\tau}, \quad t_{j,x} \in \overrightarrow{q}_i \tag{5}$$

where k is the number of the terms satisfied by resource j in query $\overrightarrow{q}_i$, m is the total number of terms in the query and τ is a parameter used to adjust the effect of the number of relevant tags in a resource profile for a query.

4.2 User Interest Relevance Measurement

User Interest Relevance Degree depends on to what extent a resource matching a user's interest. We define a function $\theta(\overrightarrow{\Psi}_i, \overrightarrow{R}_c)$ returns the User Interest Relevance Degree between user i and resource c.

$$\theta : \Psi \times R \rightarrow [0, 1]$$

where Ψ is the set of users, R is the set of resources. Different from tag-based user profile, we do not just simply calculate the similarity between resource profile and user profile to obtain user interest relevance. In our proposed tag-group based method, we firstly calculate the matching degree (denoted as ζ) of each tag-group and a specific resource profile, and then aggregate all the matching degree of tag-groups for the resource. $\zeta(\overrightarrow{g}_x, \overrightarrow{R}_c)$ returns the **tag-group match degree** between tag-group x and resource c. Here we carry out three alternative methods. The first one is "partially match", the second one is "strict match" and the third one is "binary match" (for convenience we denote them by $\breve{\zeta}$, $\acute{\zeta}$ and $\grave{\zeta}$ respectively), i.e.,

$$\breve{\zeta}(\overrightarrow{g}_x, \overrightarrow{R}_c) = \frac{\sum_{t_x \in \overrightarrow{g}_x} \omega_x}{l} \cdot \left(\frac{k}{l}\right)^{\beta} \tag{6}$$

$$\acute{\zeta}(\overrightarrow{g}_x, \overrightarrow{R}_c) = \begin{cases} \dfrac{\sum_{t_x \in \overrightarrow{g}_x} \omega_x}{l} \cdot \left(\dfrac{k}{l}\right)^{\beta} & k = n \\ 0 & k < n \end{cases} \tag{7}$$

$$\grave{\zeta}(\overrightarrow{g}_x, R_c) = \begin{cases} 1 & k = n \\ 0 & k < n \end{cases} \tag{8}$$

where k is the number of common tags in the tag-group and the resource profile, l is the number of tags in the resource profile, ω_x is the value of t_x in the resource profile, n is the number of tags in the tag-group. Intuitively "strict match" is more reasonable. We conduct experiment to compare these three ζ functions and the result verifies our expectation. Based on ζ function given above, θ function can be defined as follows.

$$\theta(\overrightarrow{\Psi}_i, \overrightarrow{R}_c) = \sum_{\overrightarrow{g}_x \in \overrightarrow{\Psi}_i} \frac{\zeta(\overrightarrow{g}_x, \overrightarrow{R}_c) \cdot e_x}{m} \tag{9}$$

where m is the number of tag-groups in $\overrightarrow{\Psi}_i$ that $\zeta(\overrightarrow{g}_x, \overrightarrow{R}_c) > 0$.

Let's take an example to see how tag-group based user profile enhances the personalized search.

Example 2: Suppose that there is a user Bob who has used "Anime Japanese" and "Action HK" and "Scientific USA" to annotate 10, 10 and 8 movies respectively. Suppose Bob issues a Query "Disaster" and consider two resources.

$$\overrightarrow{R}_1 = (Action : 1, Janpanese : 1, Disaster : 1)$$
$$\overrightarrow{R}_2 = (Scientific : 1, USA : 1, Disaster : 1)$$

Then we calculate θ for Bob By tag-based user profile, and obtained the following result.

$$\overrightarrow{U}_{Bob} = (0.357, 0.357, 0.357, 0.357, 0.286, 0.286)$$

each dimension denotes Japanese, Anime, Action, HK, Scientific and USA respectively.

$$\theta(\overrightarrow{U}_{Bob}, \overrightarrow{R}_1) = 0.714$$
$$\theta(\overrightarrow{U}_{Bob}, \overrightarrow{R}_2) = 0.572$$
$$\theta(\overrightarrow{U}_{Bob}, \overrightarrow{R}_1) > \theta(\overrightarrow{U}_{Bob}, R_2)$$

If we adopt tag-group user profile and we can get the following user profile:

$$\overrightarrow{\Psi}_{Bob} = (0.357, 0.357, 0.286)$$

each dimension denotes (Anime, Japanese), (Action, HK) and (Scientific, USA) respectively.

$$\theta(\overrightarrow{\Psi}_{Bob}, \overrightarrow{R}_1) = 0.1785, \quad \theta(\overrightarrow{\Psi}_{Bob}, \overrightarrow{R}_2) = 0.286$$
$$\theta(\overrightarrow{\Psi}_{Bob}, \overrightarrow{R}_1) < \theta(\overrightarrow{\Psi}_{Bob}, \overrightarrow{R}_2)$$

From the given information we just know Bob likes USA scientific movies and we have no ideas whether Bob likes Japanese action movies. Thus, $\theta(\overrightarrow{\Psi}_{Bob}, \overrightarrow{R}_1) < \theta(\overrightarrow{\Psi}_{Bob}, \overrightarrow{R}_2)$ is more reasonable in this example. The reason is that tag-based user profile cannot distinguish various tag-groups but tag-group based user profile can.

5. Experiments

In this section, we conduct experiments to compare our approach (denoted by *TGB*) with other baseline methods in collaborative tagging systems.

5.1 Data Set

In our experiments, we use MovieLens data set. This data set has 44805 user-item-tag-rating tuples, annotated by 2025 users on 4796 movies. Each tag is typically a single word, or a short phrase. The semantic and the purpose of particular tag are determined by each user. We randomly split the data set into test and training sets. In each group, 80 percent tuples are the training set and 20 percent tuples are the test set. Such a data splitting is to follow the splitting of most the-state-of-the-art methods (e.g., [17], [19] and [8]). We use the data in the training set to construct user profiles and resource profiles. Based on the constructed profiles, we use the data in the test set as input queries to test the effectiveness of the personalized search methods.

5.2 Evaluation Metrics

We employ three metrics here to evaluate the efficiency of our method. The first metric is *Mean reciprocal rank* (MRR) which is an overall statistical value for evaluating a ranking to a query. The reciprocal rank of a query result is the multiplicative inverse of the rank of the first correct answer. The mean reciprocal rank is the average of the reciprocal ranks of results for a query. It is defined as follows:

$$MRR = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{rank_i} \tag{10}$$

where m is the number of queries, $rank_i$ is the position of the correct answer (relevant resource) in the result ranking for the query i. The larger the average *MRR* is, the faster and easier for a personalized search method assisting the user to find out the resources he or she wants.

The second metric we use *hitrate* (HR), which is used to measure how often user interested resources are in the recommendation or personalized search result list. It is defined as:

$$HR(u_i) = \frac{|T_{u_i} \cap X_{u_i}|}{T_{u_i}} \tag{11}$$

where T_{u_i} denotes the resources relevant to (interested by) user i in the test set, and X_{u_i} is the result set of top-N returned resources. The overall hitrate of the top-N results is computed as "average HR(u_i)" for all the users in the test data set, as follows:

$$HR = \frac{\sum_{i=1}^{n} HR(u_i)}{n} \tag{12}$$

where n is the number of users. The larger the average hitrate is, the more precise the personalized search model is. The third one is *imp*. This is a common evaluation metric to measure how a personalized strategy improves the ranking of the target resources of a user in the result list by comparing to baseline methods. It is defined as:

$$imp(q_i) = \frac{1}{r_p} - \frac{1}{r_b} \tag{13}$$

where q_i is an issued query, r_b is the rank of target resource by a baseline search approach, and r_p is the rank of the same resource returned by our personalized search. The overall *ranking improvement* is calculated as "average query *imp*" for all queries in the test data, as follows:

$$imp = \frac{\sum_{i=1}^{m} imp(q_i)}{m} \tag{14}$$

where m is the number of queries. A larger value of *imp* indicates a greater improvement of the ranking for target resource by the personalization approach.

5.3 Baseline Methods

To evaluate the effectiveness of our method, we have designed two phases in the experiment. In the first step, we conduct internal comparisons on different *TGB* implementations based on three ζ functions (i.e., Eqs. (6), (7) and (8)) and two preference degree calculation methods (NTF and LOG functions) in order to find out the best configuration in the experimental environment. In the second step we compare our approach with four state-of-the-art personalized search methods in collaborative tagging system. The first one (denoted by *SIGIR '08*) is the method presented in [8], with the weights of tags in user profiles and resource profiles being based on **TF-IDF** values. The second method (denoted by *ECIR '10*) is a personalized search method from [16], in which the weights of tags in user profiles and resource profiles are an aggregation of **BM25** values and TF-IDF values. The third method (denoted by *CIKM '10*) is proposed by Cai and Li [17]. They propose a NTF method to model user profiles and resource profiles. The fourth method (denoted by *WI '12*), which is proposed by Han et al. [18], adopts a multiple-level user profile to model a user. These four methods are the current mainstream techniques for handling personalized search in collaborative tagging systems, and they use different paradigms to construct user

and resource profiles.

We conduct 10-fold cross validation and the standard deviation is small. For example, for the method *TGB@TL* which is the method we adopt to compare with the-state-of-the-art methods, its standard deviation is only 0.002 on MRR and 0.009 on Hitrate. It demonstrates that the experimental results of our propose method are stable. In our experiments, the parameter setting is the same as that claimed in baseline methods. We adopt the best setting claimed in baseline methods in our comparison experiments.

5.4 Experimental Results

Firstly we conduct experiment to find out the most suitable tag-group match function ζ and the preference degree calculation methods (NTF or LOG) in the experimental environment, which are the key steps of personalized search based on tag-group user profile. And then we compare *TGB* with the baseline methods in terms of **MRR**, **Hitrate** and **IMP**.

5.4.1 Results of Step 1

We use three different ζ functions $\check{\zeta}$, $\acute{\zeta}$ and $\grave{\zeta}$ (i.e., Eqs. (6), (7) and (8) respectively) and two preference degree calculation methods functions *NTF* and *Log* (Eqs. (1) and (2) respectively) to obtain the User Interest Relevance Degree θ (by Eq. (9)) respectively. For simple notation, we use *TGB@xy* to denote the different combinational methods where $x = P, T, A$ (using Eqs. (6), (7) and (8) respectively) and $y = N, L$ (using Eqs. (1) and (2) respectively). Their results of MRR and HR with different δ values is shown in Figs. 1 and 2.

Figures 1 and 2 shows that all the MRR and HR values of *TGB@TL* are slightly higher than *TGB@PL* and reaching the highest position with all δ. Note that almost all *TGB@* methods obtain the highest MRR and HR while δ is approaching the value 0.9. There are mainly two reasons. The reason is that using a suitable δ value to combine γ and θ is better than only using θ or γ. In other words, δ does enhance the efficiency of personalized search. Besides, using *Log* function to calculate the value for a tag-group in the user profile is better than NTF. Also, using partially match and strict match ζ functions are better than binary match in the experiment data set, which verify our observations.

Figure 3 shows the comparison of different *TGB@* methods on Hitrate. *TGB@TL* achieve the best performance in this metric as well.

In order to clearly demonstrate the effect of "tag-group based profile", we conduct experiments to compare TGB with a method based on "tag based profile", which is shown in Figs. 4 and 5. Form Figs. 4 and 5, we can find that the performance of "tag based profile with Log normalization" is better than that of "tag based profile with NTF". And our TGB method ("tag-group based profile with Log normalization" and "tag-group based profile with NTF") outperforms both of them ("tag based profile with Log normalization" and "tag based profile with NTF"). Thus, it demonstrates

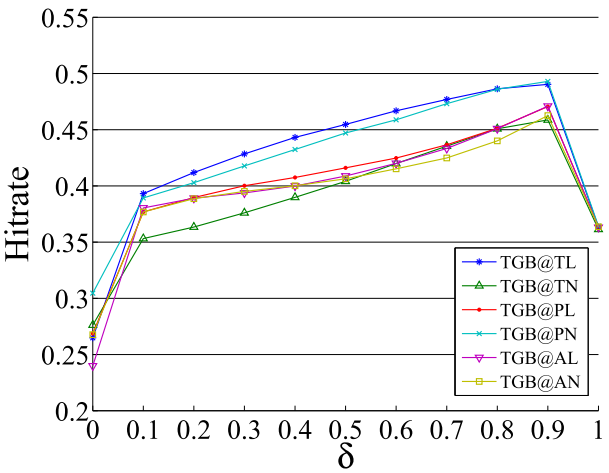


Fig. 2 Comparison of TGB@ on Hitrate@50 using MovieLens data set ($\beta = 2$, $\tau = 2$) with different δ value.

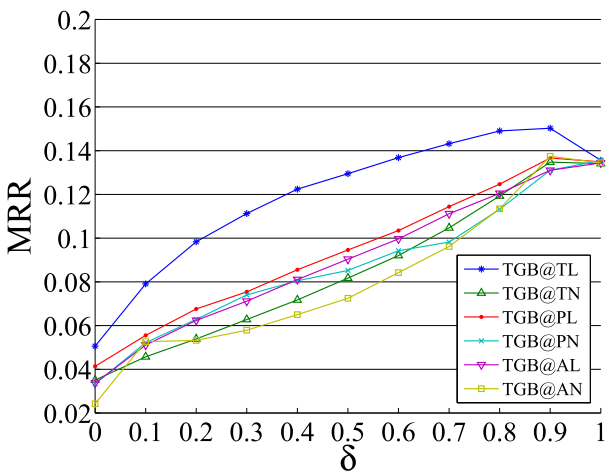


Fig. 1 Comparison of TGB@ on MRR using MovieLens data set ($\beta = 2$, $\tau = 2$) with different δ value.

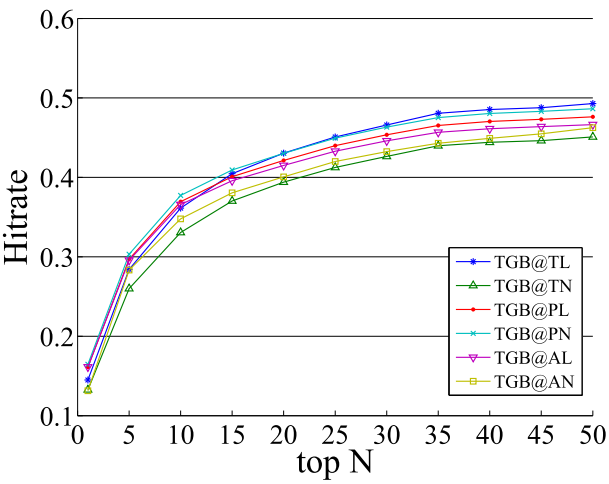


Fig. 3 Comparison of TGB@ on Hitrate@N using MovieLens data set ($\delta = 0.9$).

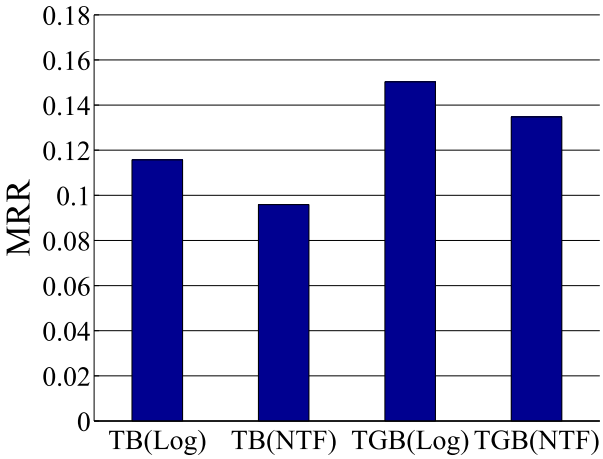


Fig. 4 Comparison of TGB method and "tag based profile" method on MRR using MovieLens data set ($\delta = 0.9$).

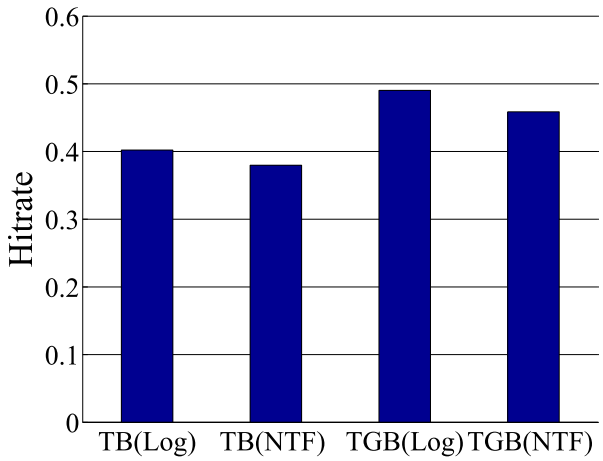


Fig. 5 Comparison of TGB method and "tag based profile" method on Hitrate@50 using MovieLens data set ($\delta = 0.9$).

that the combination of "tag-group based profile with Log normalization" is the best one.

5.4.2 Results of Step 2

To illustrate to what extent our proposed *TGB* method enhances the personalized search, we compare it with other 4 baseline methods which represent different user profiling methods. We choose *TGB@TL* as experiment setting to implement *TGB* since it achieves the best result among other *TGB@* methods. We use Eq. (2) to calculate e_x for each g_x in the user's profile as well as ω_x for each t_x in the resource's profile. Use Eq. (9) to calculate θ. And use Eq. (8) to obtain ζ.

From Fig. 6 we can see TGB obtained highest MRR (i.e., 0.15) among all the other 4 methods. Figure 7 shows the improvement of MRR between TGB and baseline methods using MovieLens data set. From Fig. 7 we can see, TGB outperforms other methods at least 21% (the best method WI'12 whose MRR is 0.124 in the adopted data set) on

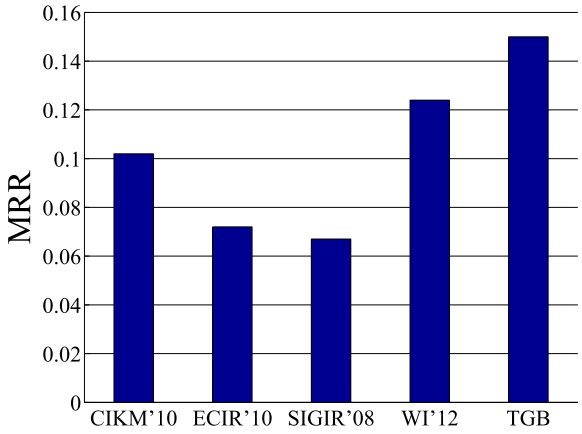


Fig. 6 Comparison of TGB and baseline methods on MRR using MovieLens data set.

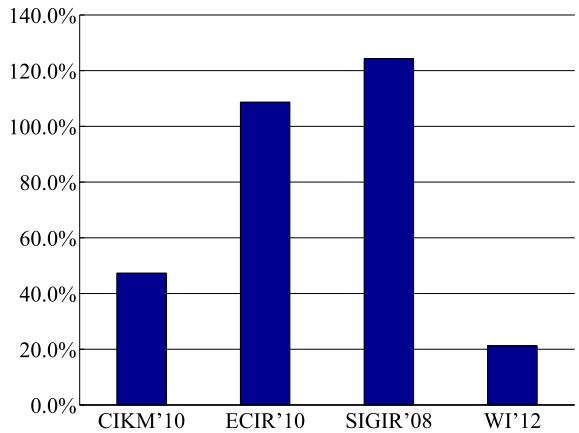


Fig. 7 Improvement of MRR between TGB and baseline methods using MovieLens data set.

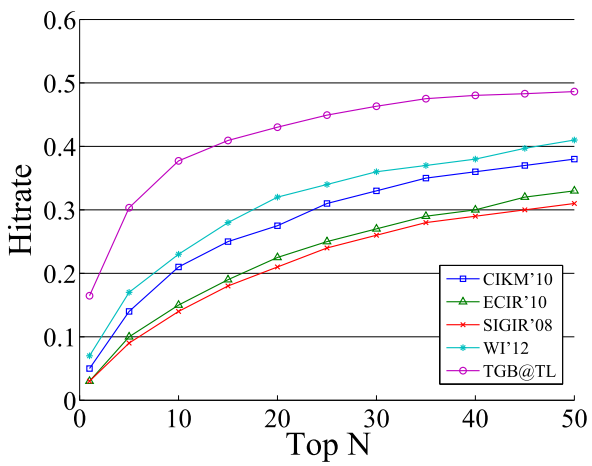


Fig. 8 Comparison of TGB and baseline methods on Hitrate using MovieLens data set.

MRR.

Figure 8 shows that TGB also outperforms the compared methods on Hitrate, especially when n is small. When $n = 1$, the Hitrate of TGB and WI'12 (the second higher Hitrate method) are 0.29 and 0.07 respectively. When $n \geq 20$,

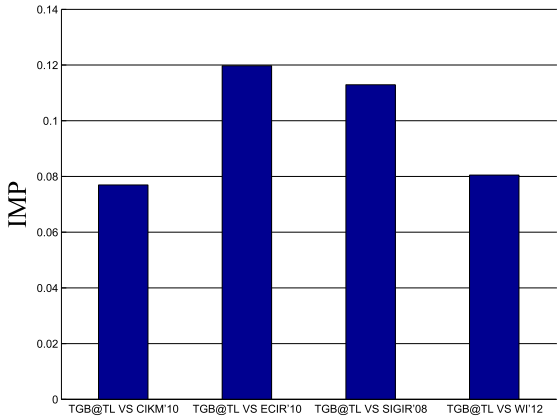


Fig. 9 Comparison of TGB and baseline methods on IMP using MovieLens data set.

the hitrate of TGB exceeds 0.5. In other word, TGB can efficiently return the correct answers for a query with a very low rank (less than 4 averagely).

Figure 9 shows the comparison of TGB and the baseline methods on IMP. We can find that TGB outperforms CIKM '10 by 7%, ECIR '10 by 12%, SIGIR '08 by 11% and WI '12 by 8%.

From Figs. 6 to 9 we can conclude that our method outperforms all the compared methods on all adopted metrics for MovieLens data set.

6. Conclusions and Future Work

In this paper, we focus on exploring the tag-group effect as well as profiling a user based on a tag-group in folksonomies. Moreover, the proposed user profile is applied to achieve personalized resource search. The core feature of tag-group based user profile is that each dimension in the user vector is a tag-group:value pair instead of tag:value pair, which can address the problem caused by tag group effect in existing profiling methods. The experimental results show that tag-group based user profile outperforms all the other baseline methods in terms of personalized search quality.

There are several potential future directions for our work. In the tag grouping process (determine the tag-groups in the user profile), we obtain tag groups according to the combination of tags annotated by a user. Since a tag-group based profile treats a set of tags as a dimension of a vector, a vector of tag-group based profile tends to be sparse. Therefore, the tag-group based method might not be work well when there is not enough training data. In our future work, we will explore to handle the data sparse problem. One possible way is to extract a subset of a whole tag-group and give different weights to subsets. Another possible way is to employ data mining methods such as association rule mining to obtain the frequent tag groups so as to expand a user profile based on the mined rules. We also plan to apply tag-based user profiling method to other applications such as recommender systems.

Acknowledgements

The research presented in this paper has been supported by National Natural Science Foundation of China (Grant No. 61300137), the Guangdong Natural Science Foundation, China (No. S2013010013836), the Fundamental Research Funds for the Central Universities, SCUT (No. 2014ZZ0035).

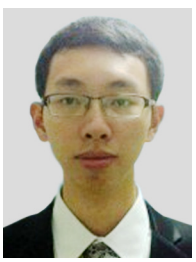
References

[1] S.A. Golder and B.A. Huberman, "Usage patterns of collaborative tagging systems," J. Inf. Sci., vol.32, no.2, pp.198–208, April 2006.
[2] K. Bischoff, C.S. Firan, W. Nejdl, and R. Paiu, "Can all tags be used for search?," Proc. 17th ACM Conference on Information and Knowledge Management, CIKM '08, pp.193–202, 2008.
[3] S. Bao, G. Xue, X. Wu, Y. Yu, B. Fei, and Z. Su, "Optimizing web search using social annotations," Proc. 16th International Conference on World Wide Web, WWW '07, pp.501–510, 2007.
[4] B. Markines, C. Cattuto, F. Menczer, D. Benz, A. Hotho, and G. Stumme, "Evaluating similarity measures for emergent semantics of social tagging," Proc. 18th International Conference on World Wide Web, WWW '09, pp.641–650, 2009.
[5] E. Michlmayr and S. Cayzer, "Learning user profiles from tagging data and leveraging them for personal(ized) information access," Proc. Workshop on Tagging and Metadata for Social Information Organization, 16th International World Wide Web Conference, 2007.
[6] H.N. Kim, I. Ha, J.G. Jung, and G.S. Jo, "User preference modeling from positive contents for personalized recommendation," Proc. 10th International Conference on Discovery Science, DS '07, pp.116–126, 2007.
[7] C.-M.A. Yeung, N. Gibbins, and N. Shadbolt, "A study of user profile generation from folksonomies," Proc. WWW Social Web and Knowledge Management, Social Web Workshop, vol.356, 2008.
[8] S. Xu, S. Bao, B. Fei, Z. Su, and Y. Yu, "Exploring folksonomy for personalized search," Proc. 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '08, pp.155–162, 2008.
[9] J. Pitkow, H. Schutze, T. Cass, R. Cooley, D. Turnbull, A. Edmonds, E. Adar, and T. Breuel, "Personalized search: A contextual computing approach may prove a breakthrough in personalized search efficiency," Commun. ACM, vol.45, no.9, pp.50–55, 2002.
[10] P.A. Chirita, C.S. Firan, and W. Nejdl, "Personalized query expansion for the web," Proc. 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '07, pp.7–14, 2007.
[11] P.A. Chirita, W. Nejdl, R. Paiu, and C. Kohlschütter, "Using ODP metadata to personalize search," Proc. 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '05, pp.178–185, 2005.
[12] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web," Proc. 7th International World Wide Web Conference, pp.161–172, 1998.
[13] K. Sugiyama, K. Hatano, and M. Yoshikawa, "Adaptive web search based on user profile constructed without any effort from users," Proc. 13th International Conference on World Wide Web, WWW '04, pp.675–684, 2004.
[14] Z. Dou, R. Song, and J.R. Wen, "A large-scale evaluation and analysis of personalized search strategies," Proc. 16th International Conference on World Wide Web, WWW '07, pp.581–590, 2007.
[15] M.G. Noll and C. Meinel, "Web search personalization via social bookmarking and tagging," ISWC '07/ASWC '07: Proc. 6th International The Semantic Web and 2nd Asian Conference on Asian Semantic Web Conference, pp.367–380, 2007.

[16] D. Vallet, I. Cantador, and J.M. Jose, "Personalizing web search with folksonomy-based user and document profiles," Advances in Information Retrieval, Proc. 32nd European Conference on IR Research, ECIR 2010, pp.420–431, 2010.
[17] Y. Cai and Q. Li, "Personalized search by tag-based user profile and resource profile in collaborative tagging systems," Proc. CIKM '10, CIKM '10, pp.969–978, 2010.
[18] H. Han, Y. Cai, Y. Shao, and Q. Li, "Improving recommendation based on features' co-occurrence effects in collaborative tagging systems," Web Technologies and Applications, pp.652–659, 2012.
[19] Y. Cai, H. Han, J. Chen, Y. Shao, H.F. Leung, and H. Min, "Integrating tags and ratings into user profiling for personalized search in collaborative tagging systems," 2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology (WI-IAT), pp.716–723, 2012.

Qing Du is currently an assistant professor of the School of Software Engineering, South China University of Technology, China. She received her B.S. degrees in Computer Applications, from South China University of Technology, China, in 2005. Her researches focus on developing effective and efficient data analysis techniques for complex data and the related applications, Intelligent software and business intelligence.

Yu Liu is currently a research assistant of School of Software Engineering, South China University of Technology, China.



Dongping Huang is currently a research assistant of School of Software Engineering, South China University of Technology, China.

Haoran Xie received the B.Eng. degree in software engineering from Beijing University of Technology, China, and the M.Sc. and Ph.D. degrees in computer science from the City University of Hong Kong, Kowloon. He is currently a senior research assistant at the Hong Kong Baptist University. His research interests include user modeling, personalization, social media, recommender systems and financial data mining.

Yi Cai received the Ph.D. degree in computer science from The Chinese University of Hong Kong. He is currently an associate professor of School of Software Engineering at the South China University of Technology, Guangzhou, China. His research interests are recommendation system, personalized search, semantic web and data mining.

Huaqing Min is a Professor and the dean of School of Software Engineering, South China University of Technology, China. His research interests includes artificial intelligence, machine learning, database, data mining and robotics.