Collaborative Filtering Recommender Systems based on Popular Tags

Huizhi Liang *Yue Xu* *Yuefeng Li* *Richi Nayak*

School of Information Technology
Queensland University of Technology
Queensland, QLD 4001, Australia

oklianghuizi@gmail.com *yue.xu@qut.edu.au* *y2.li@qut.edu.au* *r.nayak@qut.edu.au*

Abstract *The social tags in web 2.0 are becoming another important information source to profile users' interests and preferences for making personalized recommendations. However, the uncontrolled vocabulary causes a lot of problems to profile users accurately, such as ambiguity, synonyms, misspelling, low information sharing etc. To solve these problems, this paper proposes to use popular tags to represent the actual topics of tags, the content of items, and also the topic interests of users. A novel user profiling approach is proposed in this paper that first identifies popular tags, then represents users' original tags using the popular tags, finally generates users' topic interests based on the popular tags. A collaborative filtering based recommender system has been developed that builds the user profile using the proposed approach. The user profile generated using the proposed approach can represent user interests more accurately and the information sharing among users in the profile is also increased. Consequently the neighborhood of a user, which plays a crucial role in collaborative filtering based recommenders, can be much more accurately determined. The experimental results based on real world data obtained from Amazon.com show that the proposed approach outperforms other approaches.*

Keywords Information Retrieval, recommender systems, social tags, web 2.0

1 Introduction

Collaborative tagging is a new means to organize and share information resources or items on the web such as web pages, books, music tracks, people and academic papers etc. Due to the simplicity, effectiveness and being independent of the contents of items, social tags have been used in various web applications including social web page bookmarking site del.icio.us, academic paper sharing website CiteULike, and electronic commerce website Amazon.com.

A social tag is a piece of brief textural information given by users explicitly and proactively to describe and group items, thus it implies user's interests or preferences information. Therefore, the social tag information can be used to profile user's interested and preferred topics to improve personalized searching [1], generate user and item clusters [2], and make personalized recommendations [3] etc. However, as the tag terms are chosen by users freely (i.e., uncontrolled vocabularies), social tags suffer from many difficulties such as ambiguity in the meaning of and differences between terms, a proliferation of synonyms, varying levels of specificity, meaningless symbols, and lack of guidance on syntax and slight variations of spelling and phrasing [4]. These problems cause inaccurate user profiling and low information sharing among users, and also bring challenges to generate proper neighborhood for making item recommendations and consequently result in low recommendation performances. Therefore, a crucial problem in applying user tagging information to user profiling is to represent the semantic meanings of the tags.

Popular tags refer to the tags that are used by many users to collect items. Those popular tags are factual tags [5] that often capture the tagged items' content related information or topics while those tags that have low popularity are often irrelevant to the content of the tagged items or meaningless to other users, or even misspelled [5]. For one item, the popularity of using a tag to classify the item reflects the degree of common understanding to the tag and the item. High popularity means that the majority of the users think this item can be described by the tag. Thus, the popular tags reflect the common viewpoint of users or the "wisdom of crowds" [6] in the classification or descriptions of this item. Therefore, we argue that the popular tags can be used to describe the topics of the tagged items. For each user, the original tags and the collected items represent the user's personal viewpoint of item classifications and collections. In a tag, a set of items are grouped together according to the user's viewpoint. The actual topics of the tag can be described by the frequent topics of the collected items.

Proceedings of the 14th Australasian Document Computing
Symposium, Sydney, Australia, 4 December 2009.

As we just mentioned above, the major topics of each item can be represented by its popular tags, thus the popular tags of the collected items in a tag can be used to represent that tag's actual topics. Since the user's personal viewpoint of the classifications of the collected items are still kept while the original tag terms are converted to popular tags that shared by many users, the user information sharing will be improved.

In this paper, we propose to use popular tags to represent the topics of items, tags, and users' interests to solve the problems of inaccurate user profiling and low information sharing caused by the free-style vocabularies of social tags. In Section 2, the related work will be briefly reviewed. Then, the proposed collaborative filtering recommendation approach based on popular social tags will be discussed in details in Section 3. In this section, the definitions and the selection of popular social tags will be discussed firstly. Then, the approaches of representing items and tags with popular social tags will be presented. Followed by the user profiling, neighborhood formation, and recommendation generation approaches, the experimental results and evaluations will be discussed in Section 4. Finally, the conclusions will be given in Section 5.

2 Related Work

Recommender systems have been an active research area for more than a decade, and many different techniques and systems with distinct strength have been developed. Recommender systems can be broadly classified into three categories: content-based recommender systems, collaborative filtering or social filtering based recommender systems and hybrid recommender systems [7]. Because of the advantages of using similar users' recommendation and independent with the contents of items, the collaborative filtering based recommender systems have been widely used. Typically, users' explicit numeric ratings towards items are used to represent users' interests and preferences to find similar users or similar content items to make recommendations. However, because users' explicit rating information is not always available, the recommendation techniques based on user's implicit ratings have drawn more and more attention recently.

Besides the web log analysis of users' usage information such as click stream, browse history and purchase record etc., users' textural information such as tags, blogs, reviews in web 2.0 becomes an important implicit rating information source to profile users' interests and preferences to make recommendations [10]. Currently, the researches about tags in recommender systems are mainly focused on how to recommend tags to users such as using the co-occurrence of tags [2] and association rules [10] etc. Not so much work has been done on the item recommendation. Although there are some recent work which discusses about integrating tag information with content based recommender systems [11], extending the user-item matrix to user-item-tag matrix to make collaborative filtering item recommendation [12], combining uses' explicit rating with the predicted users' preferences for items based on their inferred preferences for tags [16] etc, more advanced approaches of how to exploit tags to improve the performances of item recommendations are still in demand.

More recently, the semantic meaning of social tags has become one important research question. The research of Sen etc. [5] suggests that the factual tags are more likely to be reused by different users. The work of Suchanek etc. [15] shows that popular tags are more semantically meaningful than unpopular tags. And, the research of Bischoff etc. [4] shows that not all tags are useful for searching and those tags related to the content information of items are more useful. These findings support this research. To solve the difficulties caused by the uncontrolled vocabularies of social tags, some approaches have been discussed to get the actual semantics of tags such as combining the content keywords with tags [10], using dictionaries to annotate tags [6], and contextualizing tags [17] etc. Different from these approaches, this paper proposes to use popular tags generated from the collected items to represent the semantic meanings of tags.

3 The Proposed Approach

3.1 Definitions

To describe the proposed approach, we define some key concepts and entities used in this paper as below. In this paper, tags and social tags are interchangeably used.

- **Users:** $U = \{u_1, u_2, \dots, u_n\}$ contains all users in an online community who have used tags to organize items.
- **Items or (Products, Resources):** $P = \{p_1, p_2, \dots, p_m\}$ contains all items tagged by users in *U*. Items could be any type of online information resources or products in an online community such as web pages, videos, music tracks, photos, academic papers, books etc. Each item *p* can be described by a set of tags contributed by different users.
- **Topics:** contain items' content related information such as content topics, genres, locations, attributes. For example, "globalization" is a topic that describes items' content information, "comedy" is a topic that describes items' genre information, and "Shakespeare" is a topic that describes the attribute of author information.
- **Social Tags:** $T = \{t_1, t_2, \dots, t_l\}$ contains all tags used by the users in *U*.
- **Popular social tags:** $C = \{c_1, c_2, \dots, c_q\}$ contains a set of popular social tags. Popular social tags are

tags that are used by at least θ users, where θ is a threshold. The selection of popular social tags is discussed in the followed Section.

3.2 The Selection of Popular Social Tags

Through tagging, the users, items and tags form a three dimensional relationship [12]. Based on tags, items are aggregated together if they are collected under the same tag by different users and also users are grouped together if they have used the same tag. Usually, the global popularity of a tag can be measured by the number of users that have used this tag.

Let $u(t_i)$ be the set of users who have used the tag $t_i \epsilon T$, $u(t_i, p_j)$ be the set of users who have used t_i for the item $p_j \epsilon P$, $u(t_i) = \{u(t_i, p_j) | p_j \epsilon P(t_i)\}$, where $P(t_i)$ is the set of items collected under tag t_i and $P(t_i) \subseteq P$. The global popularity of t_i can be measured by $|u(t_i)|$ which is the number of users that have used tag t_i, and the local popularity of t_i for the item p_j can be measured by $|u(t_i, p_j)|$. If we choose popular tags only based on the global popularity, some important tags that have high local popularities but relatively low global popularities (i.e., the tags that only have one kind of meaning and are used by a small number of users for tagging some particular items) will be missed out. Moreover, because a tag can have multiple meanings and users may have different understandings to the tags, some tags will have high global popularities but low local popularities such as subjective tags (i.e., "funny"). But because of the high global popularity, those tags will be incorrectly selected.

To select those popular tags that can well represent the item topics, we define the global popularity of a tag based on its maximum local popularity. Let $O(t_i)$ be the global popularity of the tag t_i, $O(t_i) = \max_{p_j \epsilon P(t_i)}\{|u(t_i, p_j)|\}$. Thus, let θ be a threshold, any tag t_i with $O(t_i) > \theta$ will be selected as a popular social tag.

Theoretically, the threshold θ can be any positive numbers. However, since $O(t_i)$ is the maximum local popularity of t_i for its collected items, if θ is too large, the number of popular tags will be small, and there might be some items which are not tagged by any of those selected popular tags. On the other hand, each item collects a set of tags that have been used by different users to tag this item. Let $T(p_j)$ be the collected tag set of p_j, $\max_{t_i \epsilon T(p_j)}\{|u(t_i, p_j)|\}$ is the maximum local popularity of the tags in $T(p_j)$ for item p_j. Apparently, if $\theta > \max_{t_i \epsilon T(p_j)}\{|u(t_i, p_j)|\}$, then all the tags of item p_j will be excluded which will result in no popular tags to describe the topics of p_j. To avoid this situation, we define an upper boundary for the threshold θ. Let $\lambda = min_{p_j \epsilon P}\{\max_{t_i \epsilon T(p_j)}\{|u(t_i, p_j)|\}\}$. If $\theta \leq \lambda$, then each item can be guaranteed to have at least one popular tag to describe it. Therefore, the popular social tag set C also can be denoted as:

$$C = \{t_i | O(t_i) \geq \theta, t_i \in T, \lambda \geq \theta > 0\}, C \subseteq T.$$

3.3 Item and Tag Representations

The selected popular tags are used to represent items' major topics and the actual topics of each user's tags.

Item Representation

Traditionally, the item classifications or descriptions are given by experts using a set of standard and controlled vocabulary as well as a hierarchical structure representing the semantic relationships among the topics to describe the topics of the items such as item taxonomy and ontology. In web 2.0, harnessing the collaborative work of thousands or millions of web users, the aggregated tags contributed by different users form the item classifications or descriptions from the viewpoint of users or folksonomy [13]. For each item p_j, the set of tags used by users to tag p_j, denoted as $T(p_j)$, and the number of users for each tag in $T(p_j)$ form the item description of item p_j, which is defined as below.

Definition 1 (Item Description): Let p_j be an item, the item description of p_j is defined as the set of social tags for p_j and their numbers of being used to tag the item p_j, which is denoted as $D(p_j) = \{(t_i, O(t_i, p_j)) | t_i \in T(p_j), O(t_i, p_j) > 0\}$, where $O(t_i, p_j)$ is the number of users that use the tag t_i to tag the item p_j and $O(t_i, p_j) = |u(t_i, p_j)|$.

An example of item description is shown in Figure 1. The book "*The World is Flat"* is described by 10 tags such as "globalization", "economics", "business" etc. and their user numbers.

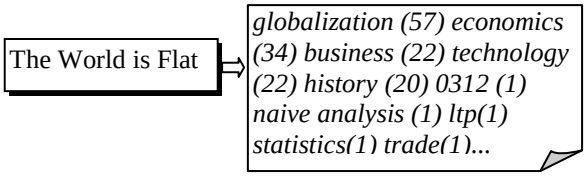

Figure 1: An example of item description formed by social tags.

Different from the item descriptions or classifications provided by experts, the item descriptions formed by social tags contain a lot of noise, which brings challenges for the organizing, sharing and retrieval of items. However, an advantage provided by the item descriptions formed by social tags is that the item description $D(p_j)$ records the user number of each tag for p_j or the local popularity of each tag for p_j. This feature can be used to find the major topics of items and filter out the noise. For example, in Figure 1, we can see that 57 users use the tag "globalization" to classify the book "The World is Flat", which is the most frequently used tag to tag this book, and the term "globalization" is indeed the actual

major topic of this book. Moreover, the tag "0312" only has one user, and it doesn't reveal any information in terms of the topics of the book. Removing the unpopular tags such as "0312" won't reduce the coverage of the remaining tags to represent the topics of the book but the noise. Therefore, we propose to use the selected popular tags to represent the items.

Definition 2 (Item Representation) Let p_j be an item, $C = \{c_1, c_2, \dots, c_q\}$ be the set of popular tags, the representation of p_j is defined as a set of popular social tags along with their frequencies as described below:
$IR(p_j) = \{(c_x, f(p_j, c_x)) | c_x \in C, f(p_j, c_x) > 0\}$,
$f(p_j, c_x) = O(c_x, p_j) / \sum_{c_y \in C} O(c_y, p_j)$, where $f(p_j, c_x)$ is the frequency of c_x for p_j, $f(p_j, c_x) \in [0,1]$ and $\sum_{c_x \in C} f(p_j, c_x) = 1$.

The frequency $f(p_j, c_x)$ represents the degree of item p_j belonging to c_x. For a given set of popular tags C with size q, i.e.,$|C| = q$,the topics of each item $p_j \in P$ can be represented by a vector $\overrightarrow{b_j} = (b_{j,1}, b_{j,2}, \dots, b_{j,x}, \dots, b_{j,|C|})$, where $b_{j,x} = f(p_j, c_x)$. Thus, for each item p_j, its topic representation becomes:
$\overrightarrow{b_j} = (b_{j,1}, b_{j,2}, \dots, b_{j,x}, \dots, b_{j,|C|})$

Tag Representation

As mentioned in Introduction, since the unrestricted nature of tagging, social tags contain a lot of noise and suffer some problems such as semantic ambiguity and a lot of synonyms etc., which brings challenges to make use of social tags to profile users' interests accurately.

Although not all tags are meaningful to other users or can be used to represent the topics, for each user, his/her own tags and items collected with those tags reflect that user's personal viewpoint of classification of the collected items. Thus, each tag used by a user is useful for profiling that user no matter how popular this tag is. In a tag, a set of items are grouped together according to a user's viewpoint, therefore, the frequent topics of these items can be used to represent the actual topics of the tag. Since the major topics of each item can be represented by its popular tags, the frequent popular tags of the collected items in a tag can be used to represent that tag's actual covered or related topics.

Definition 3 (Tag Representation): Let t be a tag used by user u, $C = \{c_1, c_2, \dots, c_q\}$ be the set of popular tags, the representation of t is defined as a set of weighted popular social tags as described below:
$TR(t,u) = \{(c_x, w(c_x,t,u)) | c_x \in C, w(c_x,t,u)) > 0\}$, where $w(c_x,t,u)$ is the weight of c_x, $w(c_x,t,u) \in [0,1]$, $\sum_{c_x \in C} w(c_x,t,u) = 1$.

The weight of c_x or $w(c_x,t,u)$ can be measured through calculating the total frequency of c_x for all the items collected in the tag t by the user u. Since the number of items in different tags may be different, we normalize $w(c_x,t,u)$ with the number of items in the tag t of u. Let $P(t,u)$ denote the set of items that are collected or classified to the tag t by user u, then the weight of c_x can be calculated as below:
$w(c_x,t,u) = \frac{1}{|P(t,u)|} \sum_{p_j \in P(t,u)} f(p_j, c_x)$, where $f(p_j, c_x)$ is the frequency of c_x for the item p_j in the tag t, as shown in Definition 2, $f(p_j, c_x) = O(c_x, p_j) / \sum_{c_y \in C} O(c_y, p_j)$.
Apparently, the tag representation $TR(t,u)$ *is* generated based on the items collected in the tag t by the user u. That means, $TR(t,u)$ still reflects the personal viewpoint of the user u about the item classifications or collections. Thus, each user's viewpoint of classifying his/her items is still kept while a set of popular tags are obtained to represent each tag term's semantic meaning. For different users, the representations for the same tag can be different. On the other hand, for different users, the representations for different tags can be the same or similar. Even though the tag terms are freely chosen by individual users, by representing each tag using a set of popular tags, all tags become comparable since all of them are represented using the same set of terms (i.e., popular tags). With the popular tag representation, those unpopular tags that often cause confusions and noises become understandable by other users according to the understanding to their corresponding popular tag representation. For those popular tags, their tag representations reveal other related popular tags, very often, these popular tags themselves have high weight in their tag representation. Since each tag is represented by a set of popular tags which provides the ground for comparison, this approach can help to solve the problems caused by the free style vocabulary of tags such as tag synonyms which means some different tags have the same meaning, semantic ambiguity of tags which means one tag has different meanings for different users, and spelling variations etc.

3.3 User Profile Generation

User profile is used to describe user's interests and preferences information. Usually, a user-item rating matrix is used in collaborative filtering based recommender systems to profile users' interests, which are used to find similar users through calculating the similarity of item ratings or the overlaps of item sets [14]. With the tag information, users can be described with the matrix (user, (tag, item)), where (tag, item) is a sub matrix representing the relationship between the tag set and item set of each user. Binary values "1" and "0" are used to specify whether a tag or an item has been used or tagged by a user or not. Through calculating the overlaps of tags and items or each user's sub relationship of tags and items, neighborhood can be

formed to do collaborative filtering to recommend items to a target user [12][3].

As mentioned before, the free-style vocabulary of tags causes a lot of noise in tags which resulted in inaccurate user profiles and incorrect neighbors. Moreover, because of the long tails of items and tags, the size of the matrix is very big and the overlaps of commonly used tags and tagged items are very low, which makes it difficult to find similar users through calculating the overlaps of tags and items. To solve these problems, we propose to profile users' interests to topics by using a set of popular tags and convert the binary matrix (user, (tag, item)) into a much smaller sized user-topics matrix. The popular tags will be used to represent each user's interested topics and numeric scores will be used to represent how much the user are interested in these topics.

Definition 4 (User Profile): Let u_i be a user, $C = \{c_1, c_2, \dots, c_q\}$ be the set of popular tags, the user profile of u_i is defined as a $|C|$-sized vector with scores reflecting user's interests to the popular tags, which is donated as $\overrightarrow{v_i} = (v_{i,1}, v_{i,2}, \dots, v_{i,x}, \dots, v_{i,|C|}) = \big(sc(u_i, c_1), sc(u_i, c_2), \dots, sc(u_i, c_x), \dots, sc(u_i, c_q)\big)$. $sc(u_i, c_x)$ is the score to $v_{i,x}$ that represents the degree of u_i's interests to the popular tag c_x.

A matrix $\vec{v}$ with size $|U| \times |C|$, can be used to represent the user profiles for all users in U. Each row $\overrightarrow{v_i}$ in the matrix $\vec{v}$ represents the user profile of user u_i. In order to facilitate the similarity measure of any two users, user-wise normalization is applied. We suppose each $u_i \epsilon\, U$ has the same total interest score N and $\sum_{c_x \in C} sc(u_i, c_x) = N$, where N is the normalization factor, which can be any positive number. Thus, $sc(u_i, c_x) \in [0, N]$.

To calculate each user's topic interest degree $sc(u_j, c_x)$, firstly, we calculate the user's interest distribution for his/her own original tags. Let $T_i = \{t_{i,1}, t_{i,k}, \dots, t_{i,a}\}$ be the tag set of u_i, $t_{i,1}, t_{i,k}, \dots, t_{i,a} \epsilon T$, $s(t_{i,k})$ be the score to measure how much u_i is interested in $t_{i,k}$, then the score vector $(s(t_{i,1}), s(t_{i,k}), \dots, s(t_{i,a}))$ will represent u_i's interest distribution over his/her own tags, $\sum_{k=1}^{a} s(t_{i,k}) = N$.

A common sense is that, if a user is more interested in a tag or topic, usually the user may collect more items under that tag or about that topic. That means, the number of items in a tag is an important indicator about how much the user is interested in the tag. Let $|P(t_{i,k}, u_i)|$ denote the number of items in the tag $t_{i,k}$ used by user u_i, we use the proportion of $|P(t_{i,k}, u_i)|$ to the total number of items in all tags of u_i to measure the user's interest degree to the tag $t_{i,k}$. Thus, $sc(t_{i,k})$ can be calculated as shown as follows:

$$s(t_{i,k}) = N \cdot \frac{|P(t_{i,k}, u_i)|}{\sum_{k=1}^{a} |P(t_{i,k}, u_i)|} \qquad (1)$$

By using Equation 1, we can obtain the user-tag matrix that describes tag interests of all the users. As discussed before, a tag can be represented with a set of popular social tags derived from the collected items with that tag. We can calculate the score of user u_i to topic c_x in each tag $t_{i,k}$ denoted as $c_{x,k}$ for the user u_i, shown as below:

$$sc(u_i, c_{x,k}) = s(t_{i,k}) \cdot w(c_{x,k}, t_{i,k}, u_i), x = 1..q, k = 1..a \qquad (2)$$

The user's interest score to the topic c_x, $sc'(u_i, c_x)$, is calculated by summing up the user's interests to the topic in all his tags:

$$sc(u_i, c_x) = \sum_{k=1}^{a} sc(u_i, c_{x,k}) \qquad (3)$$

With Equation 3, users' interest distributions over their own original tags are converted to users' interest distributions over the topics of items that are represented by the popular tags. Using this user profiling approach, the noise of social tags can be greatly removed while each user's personal viewpoint of classifications or collections will still remain. Moreover, since the size of the converted matrix is much smaller than the size of the matrix (user, (tag, item)), the information sharing among different users can be improved as well.

3.4 Neighborhood Formation

Neighborhood formation is to generate a set of like-minded peers for a target user. Forming a neighborhood for a target user $u_i \epsilon$ U with standard "best-*K*-neighbors" technique involves computing the distances between u_i and all other users and selecting the top K neighbors with shortest distances to u_i. Based on user profiles, the similarity of users can be calculated through various proximity measures. Pearson correlation and cosine similarity are widely used to calculate the similarity based on numeric values.

Based on the user profiles discussed above, for any two users u_i and u_j with profile v_i and v_j, the Pearson correlation is used to calculate the similarity, which is defined as below:

$$sim(u_i, u_j) = \frac{\sum_{y=1}^{q} (v_{i,y} - \overline{v_i}) \cdot (v_{j,y} - \overline{v_j})}{\sum_{y=1}^{q} (v_{i,y} - \overline{v_i})^2 \cdot \sum_{y=1}^{q} (v_{j,y} - \overline{v_j})^2)} \qquad (4)$$

Using the similarity measure approach, we can generate the neighborhood of the target user u_i, which includes K nearest neighbour users who have similar topic interests with u_i. The neighbourhood of u_i, is denoted as:

$$\check{N}(u_i) = \{u_j | u_j \epsilon\, maxK\{sim(u_i, u_j)\}, u_j \epsilon U$$

where *maxK {}* is to get the top K values.

3.5 Recommendation Generation

For each target user u_i, a set of candidate items will be generated from the items tagged by u_i 's neighbourhood formed based on the similarity of users, which is denoted as $\check{C}(u_i)$, $\check{C}(u_i) = \{p_k | p_k \epsilon P(u_j), u_j \epsilon \check{N}(u_i), p_k \notin P(u_i)\}$, where $P(u_j)$

is the item set of user u_j . With the typical collaborative filtering approach, those items that have been collected by the nearest neighbors will be recommended to the target user.

As discussed in Section 3.2, the aggregated social tags describe the content information of items and the topics of each item can be represented by popular social tags. Thus, we propose to combine the content information of items formed by popular social tags with the typical collaborative filtering approach to generate recommendations. Those items that not only have been collected by the nearest neighbors but also have the most similar topics to the target user's interests will be recommended to the target user, which makes the proposed recommendation generation approach actually get the benefits of the content based recommendation approaches [8].

For each candidate item $p_k \epsilon \check{C}(u_i)$, let $\check{N}(u_i, p_k)$ be the set of users in $\check{N}(u_i)$ who have tagged the item p_k, the prediction score of how much u_i may be interested in p_k is calculated in terms of the aspects of how similar those users who have the item p_k and how similar the item's topics with u_i's topic interest.

With Equation 4, the similarity of two users can be measured. Similarly, the Pearson correlation is used to calculate the similarity of the topic interests of user u_i and the topics of the candidate item p_k, which is denoted as below:

$$sim(u_i, p_k) = \frac{\sum_{y=1}^{q}(v_{i,y} - \overline{v_i}) \cdot (b_{k,y} - \overline{b_k})}{\sum_{y=1}^{q}(v_{i,y} - \overline{v_i})^2 \cdot \sum_{y=1}^{q}(b_{k,y} - \overline{b_k})^2)} \quad (5)$$

Thus, the prediction score denoted as $A(u_i, p_k)$ can be calculated with Equation 6.

$$A(u_i, p_k) = \frac{sim(u_i, p_k) \cdot \sum_{u_j \epsilon \check{N}(u_i, p_k)} sim(u_i, u_j)}{|\check{N}(u_i, p_k)|} \quad (6)$$

The top *N* items with larger prediction scores will be recommended to the target user u_i.

4 Experiments and Evaluations

4.1 Experiment setup

We conducted the experiments using the dataset obtained from Amazon.com. The dataset was crawled from amazon.com on April, 2008. The items of the dataset are books. To avoid too sparse, in pre-processing, we removed the books that are only tagged by one user. The final dataset comprises 5177 users, 37120 tags, 31724 books and 242496 records.

The precision and recall are used to evaluate the recommendation performance. The whole dataset is split into a training dataset and a test dataset with 5-folded and the split percentage is 80% for the training dataset and 20% for the test dataset, respectively. Because our purpose is to recommend books to users, the test dataset only contain users' books information. Each record in the test dataset consists of the books that are tagged by one user. The training dataset, which is used to build user profiles, contains users' books and corresponding tags information as well. For each user in the test dataset, the top N items will be recommended to the user. If any item in the recommendation list is in the target user's testing set, then the item is counted as a hit.

4.2 Parameterization

The global popularities of tags are shown in Figure 2. We can see that the user number of tags follows the power law distribution, which means that a small number of tags are used by a large number of users while a large number of tags are only used by a small number of users. Among 37120 tags, there are about 67% tags (i.e., 25006 tags) which are only used by one user.

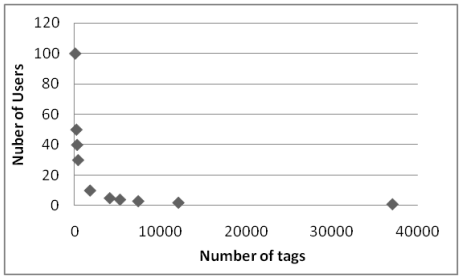

Figure 2: The distribution of social tags.

After calculating the local popularity of each tag for each item, we get λ=2. Thus, we set θ=2. To evaluate the effectiveness of the selected popular tag set, we compared the top 5 precision and recall results of the threshold θ=2 with the results of θ =1, θ =3, θ =4, and θ =5. With threshold θ =1, 37120 tags are selected, which is the whole tag set. Thus, each item was represented with all the tags. Different from the *Topic-Tag* approach, each tag was represented with the selected tags. With threshold θ =2, 12214 tags are selected. When threshold θ =3, 7428 tags were selected and there were 1188 books that have no selected tags describes them. With threshold θ =4, 5297 tags were selected and there were 1668 books that have no selected tags describes them. With threshold θ =5, 4104 tags were selected and there were 2452 books that have no selected tags describes them. The top 5 precision and recall results with different threshold are shown in Figure 3.

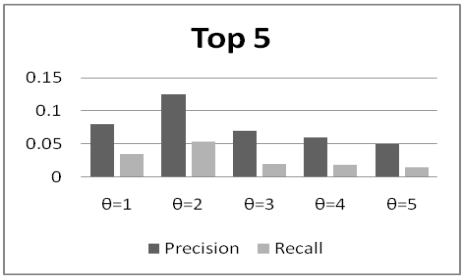

Figure 3. The top 5 precision and recall evaluation results with different threshold θ values.

From the results of Figure 3, we can see the results of θ =2 was better than other values. Thus, the popular tags can be used to represent the topics of items and tags. And, since some books may don't have any selected tags describing their topics when the threshold is too high, the results are worse.

4.3 Comparison

To evaluate the effectiveness of the proposed approach, we compared the precision and recall of the recommended top N items produced by the following approaches:

- ***Topic-PopularTag approach***. This is the proposed approach that uses the popular tag to represent items' topics, tags' actual topics and users' topic interests.
- ***Topic-Tag approach.*** This approach uses users' interest distribution to their original tags to make recommendation. Different from *Topic-PopularTag* approach, this approach only uses the users' original tags to profile users and doesn't include the tag representations.
- ***Singular Value Decomposition (SVD).*** This is a wildly used approach to reduce the dimensions of a matrix and reduce noise. In this paper, the standard SVD based recommendation approach [8] was implemented based on the user-tag matrix.
- ***Tso-Sutter's approach***. This approach is proposed by Tso-Sutter that uses two derived binary matrixes user-item, user-tag to make recommendation [9], which is an extended standard collaborative filtering approach.
- ***Liang's approach.*** This approach is proposed by Liang that uses three derived binary matrixes user-item, user-tag to tag-item sub matrix to make recommendation [12], which is an extended standard collaborative filtering approach.
- ***Standard CF approach.*** This is the standard collaborative filtering (CF) approach [14] that uses the implicit item ratings or the binary matrix user-item only. This is the baseline approach.

We compared the proposed approach that has the threshold θ =2 with other state of art approaches, the precision and recall results are shown in Figure 4 and Figure 5.

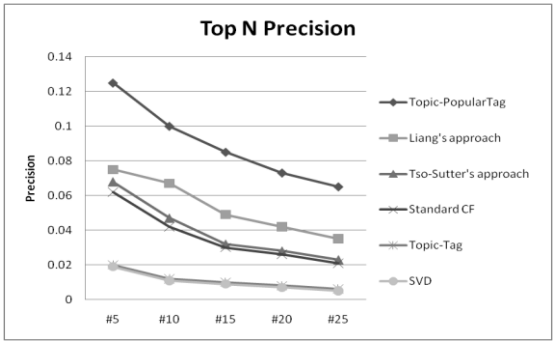

Figure 4: Precision evaluation results.

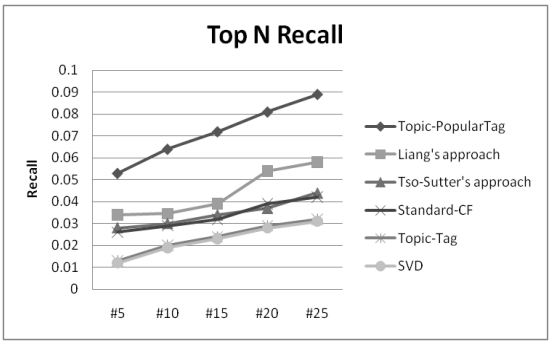

Figure 5: Recall evaluation results.

4.4 Discussions

From the experimental results, we can see that the proposed approach outperformed the other approaches, which means the proposed collaborative filtering approach based on popular social tags is effective. Since the dataset is very sparse (i.e., the average number of items that each user has is about 12.6), the overall precision and recall values are low. The approach *Topic-Tag* approach performed the worst, which means that although tags implies users' interests and preferences information, since the social tags contains a lot of noise, it's inaccurate to profile users with their original tags directly. The comparison between the approaches of *Tso-Sutter* and *Liang* and the *Standard CF* approach shows that social tags are helpful to improve the user profiling accuracy when the social tags are used together with the users' collected items. Moreover, the comparison between the proposed *Topic-PopularTag* approach and the *SVD* approach suggests that the proposed approach performs better than the traditional dimension reduction approach. The proposed approach not only reduce the dimension through using a much smaller sized user-topic matrix to profile users but also significantly improves the accuracy of user profiling and information sharing through representing the personal or unpopular tags with a set of popular tags.

5. Conclusions

In this paper, we propose a collaborative filtering approach that combines each user's personal viewpoint of the classifications of items and the common viewpoint of many users about the classifications of items to make personalized item recommendation. The popular tags are used to represent items' major topics, tags' actual covered or related topics and users' topic interests. Moreover, a user profiling approach that converts users' interest distribution for their own original tags to users' interest distribution for topics that are represented with the popular tags are proposed to improve user profiling accuracy and information sharing. Also, we propose a recommendation generation approach that incorporates the item content

information formed by the collaborative working of tagging to generate recommended items that are not only have been collected by most similar users but also have the most similar topics with the target user's interests.

The experiments show that the proposed approach outperforms other approaches. Since the social tags can be used to describe any types of items or resources, this research can be used to recommend various kinds of items to users, which provides possible solutions to the recommendation of those items that the traditional collaborative filtering approaches or content based approaches fail to work well such as people. Moreover, this research made a contribution to the improvement of information sharing, organization and retrieval of online tagging systems as well as the improvement of the recommendation performances of traditional recommender systems (i.e., in e-commerce websites) through incorporating this new type of user information in web 2.0.

References

[1] Bao, S., Wu, X., Fei, B., Xue, G., Su, Z. and Yu, Y., "Optimizing Web Search Using Social Annotations", In *Proc. of WWW'07*, 2007, pp. 501-510.

[2] Li, X., Guo, L., and Zhao, Y. E., "Tag-based social interest discovery", In *Proc. of WWW'08*, 2008, pp. 675-684.

[3] Tso-Sutter, K.H.L., Marinho, L.B. and Schmidt-Thieme, L., "Tag-aware Recommender Systems by Fusion of Collaborative Filtering Algorithms", In *Proc. of Applied Computing*, 2008, pp. 1995-1999.

[4] Bischoff, K., Firan, C. S., Nejdl, W., Paiu, R., "Can All Tags be Used for Search?", In *Proc. of CIKM'08*, 2008, pp. 193-202.

[5] Sen, S., S. Lam, A. Rashid, D. Cosley, D. Frankowski, J.Osterhouse, M. Harper, and J. Riedl., "Tagging, communities, vocabulary, evolution", In *Proc. of CSCW '06*, 2006, pp. 181-190.

[6] What Is Web 2.0. http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html

[7] Burke, R., "Hybrid Recommender Systems: Survey and Experiments", *User Modeling and User-Adapted Interaction*, 12(2002), pp. 331-370.

[8] Sarwar, B. M., Karypis, G., Konstan, J. A., and Riedl, J. "Application of Dimensionality Reduction in Recommender System — A Case Study." In *Proc.of WebKDD'00, 2000*.

[9] K.H.L. Tso-Sutter, L.B. Marinho and L.Schmidt-Thieme, "Tag-aware Recommender Systems by Fusion of Collaborative Filtering Algorithms", In *Proc. Applied Computing*'08, 2008, pp.1995-1999.

[10] Heymann, P., Ramage, D., and Garcia-Molina, H., "Social tag prediction", In *Proc. of SIGIR'08*, 2008, pp. 531–538.

[11] Gemmis, M. de, Lops, P., Semeraro, G., and Basile, P., "Integrating tags in a semantic content-based recommender", In *Proc. of the 2008 ACM conference on Recommender systems*, 2008, pp. 163-170.

[12] Liang, H., Xu, Y., Li, Y., and Nayak, R., "Collaborative Filtering Recommender Systems Using Tag Information", In *Proc. of The 2008 IEEE/WIC/ACM International Conference on Web Intelligence (WI-08) Workshops*, 2008, pp. 59-62.

[13] Al-Khalifa, H.S. and Davis, H. C., "Exploring the Value of Folksonomies for Creating Semantic Metadata", *International Journal on Semantic Web and Information Systems*, 3,1 (2007), pp. 13-39.

[14] Shardanand, U. and Maes,P., "Social Information Filtering: Algorithms for Automating 'Word of Mouth'", In *Proc. of SIGCHI*, 1995, pp. 210 -217.

[15] Suchanek, F. M., Vojnovi ć, M., Gunawardena D., "Social tags: Meaning and Suggestions", In *Proc.of CIKM'08,* 2008, pp. 223-232

[16] Sen, S., Vig, J., Riedl, J., "Tagommenders: Connecting Users to Items through Tags", In *Proc. of WWW'09*, 2009, pp. 671-680

[17] Au Yeung, C. M., Gibbins, N. and Shadbolt, N., "Contextualizing Tags in Collaborative Tagging Systems", In *Proc. of the 20th ACM Conference on Hypertext and Hypermedia*, 2009.