Tag Based Collaborative Filtering for Recommender Systems

Huizhi Liang, Yue Xu, Yuefeng Li, and Richi Nayak

School of Information Technology, Queensland University of Technology,
Brisbane, Australia
oklianghuizi@gmail.com, {yue.xu,y2.li,r.nayak}@qut.edu.au

Abstract. Collaborative tagging can help users organize, share and retrieve information in an easy and quick way. For the collaborative tagging information implies user's important personal preference information, it can be used to recommend personalized items to users. This paper proposes a novel tag-based collaborative filtering approach for recommending personalized items to users of online communities that are equipped with tagging facilities. Based on the distinctive three dimensional relationships among users, tags and items, a new similarity measure method is proposed to generate the neighborhood of users with similar tagging behavior instead of similar implicit ratings. The promising experiment result shows that by using the tagging information the proposed approach outperforms the standard user and item based collaborative filtering approaches.

Keywords: Collaborative filtering, collaborative tagging, recommender systems, user profiling.

1 Introduction

Nowadays collaborative tagging or social annotation is becoming popular in online web sites or online communities. Harnessing the collaborative work of thousands or millions of web users to add natural language keywords to information resources, it becomes easy to retrieve, organize and share information quickly and efficiently. For its simplicity and effectiveness, collaborative tagging has been used in various web application areas, such as social bookmark site del.ici.ous, photo sharing website Flickr.com, academic paper database system CiteULike, and electronic commerce website Amazon.com, etc.

Besides helping user organizing his or her personal collections, a tag also can be regarded as a user's personal opinion expression while tagging can be considered as implicit rating or voting on the tagged information resources or items [1]. Thus, the tagging information implies user's important personal interest and preference information, which can be used to greatly improve personalized searching [2] and recommendation making.

Currently some works have been done on how to use collaborative tagging information to recommend personalized tags to users [3], but not much work done on utilizing tagging information to help users to find interested items easily and quickly.

P. Wen et al. (Eds.): RSKT 2009, LNCS 5589, pp. 666–673, 2009.

Thus, how to recommend personalized items to users based on tagging information becomes an important research question and the research is just on the start.

In this paper, we will propose a tag-based collaborative filtering approach that can make personalized recommendations based on user tagging behavior. The paper is organized as below:

In section 2, the related work will be discussed. Then, we will discuss the proposed tag-based collaborative filtering approach in details in section 3. In this section, the user profiling approach, the distinctive three- dimensional relationship among users, items and tags, the similarity measure method and the user-based and item-based approaches of generating top N recommended item list will be discussed. The experiments will be illustrated in section 4 while the discussion about the experiment results will be on section 5. Finally, in section 6, we will give a conclusion about this work.

2 Related Work

Collaborative tagging is a typical web 2.0 application that contains plenty of user interaction information. Collaborative tagging information can be used to build virtual social network, find interest group as well as organize, share, gather and discover information resources. As collaborative tagging information is a kind of emergent online community information, the discussion of tagging behavior itself and its usage patterns and applications still remain open [4].

Collaborative filtering is a traditional and wildly used approach to recommend items to users, which based on the assumption that similar minded people will have similar taste or behaviors. Although there is a lot of work on the collaborative filtering recommender systems, only Tso-Sutter's [5] work discussed about using the tag information to do item recommendation to the best of our knowledge.

In Tso-Sutter's work, the three-dimensional relationship among user, item and tag was converted into three two dimensional relationships user-item, user-tag and tag-item. Thus, the tag information was used as an extension of user-item implicit rating matrix and the tagging behavior was profiled and measured as implicit rating behavior. Because it ignored some distinct features of tagging behavior, the work failed to use tag information to do item recommendation accurately.

3 Tag Based Collaborative Filtering

3.1 User Profiling

User profiling is to model users' features or preferences. The approaches of profiling users with user-item rating matrix and keywords vectors are widely used in recommender systems. To profile user's tagging behavior correctly and accurately, we propose to model a user in a collaborative tagging community in three aspects, i.e., the tags used by the user, the items tagged by the user, and the relationship between the tags and the tagged items. For easy describing the proposed approach, we give the following definitions:

U: Set of users. U= $\{u_1, u_2 \ldots u_n\}$, it contains all the users of the collaborative tagging community.

P: Set of items. P= $\{p_1, p_2\ldots p_m\}$, it contains all tagged items. An item is an object that is tagged by users and it can be any kind of objects in the application areas, such as books, movies, URLs, photos, and academic papers etc.

T: Set of tags. T= $\{t_1, t_2\ldots, t_l\}$, it includes all the tags that have been used by users. A tag is a relevant keyword assigned to one or more items by a user, describing the items and enabling classification of the items.

$E(u_i,t_j,p_k)$: a function that specifies user u_i used the tag t_j tagging item p_k

The user profile is defined as below:

Definition [User Profile]: For a user u_i, i=1..n, let Tu_i be the tag set of u_i, $Tu_i=\{t_j|t_j\in T, \exists p_k\in P, E(u_i,t_j, p_k) =1\}$, $Tu_i \subseteq T$, Pu_i be the item set of u_i, $Pu_i=\{p_k|p_k\in T, \exists t_j\in P, E(u_i,t_j, p_k) =1\}$, $Pu_i \subseteq P$, TP_i be the relationship between u_i's tag and item set, $TP_i=\{<t_j, p_k>| t_j\in T, p_k\in P$, and $E(u_i,t_j,p_k)=1\}$, $UF_i = (Tu_i, Pu_i, TP_i)$ is defined as the user profile of user u_i. The user profile or user model of all users is denoted as UF, $UF=\{UF_i|i=1..n \}$.

3.2 The Multiple Relationships

From the above user profile, we can see the relationship describing the situation of an item p_k being tagged with tag t_j by user u_i is three-dimensional, which is very different with the two-dimensional explicit rating behavior or other implicit rating behaviors that only involve users and items. Based on it, other three-dimensional and two-dimensional relationships can be derived. These multiple relationships are vital for collaborative filtering approaches especially for the neighborhood forming.

To facilitate understanding, we discuss the multiple relationships among users, tags and items from the perspectives of user, item and tag respectively as follows:

- From the perspective of users, the relationship among users, tags and items is denoted as $R_{U, TP}$, which is the direct and basic three-dimensional relationship and describes the tagging behavior of each user. $R_{U, TP}=\{<u_i,TP_i>|u_i\in U, i=1..n\}$, where TP_i is the relationship between u_i's tag and item set, as defined in section 3.1. Based on it, other two two-dimensional relationships $R_{U, P}$ and $R_{U, T}$ can be derived, which are defined as below:

$R_{U, P}$: The relationship between users and their item sets. This two dimensional relationship can be used as the base of traditional user-based collaborative filtering approach. $R_{U, P} = \{<u_i,Pu_i>|u_i\in U, Pu_i \subseteq P, i=1..n\}$, Pu_i is item set of u_i, as defined in section 3.1.

$R_{U, T}$: The relationship between users and their tag sets. $R_{U, T} = \{<u_i,Tu_i>|u_i\in U, Tu_i \subseteq T, i=1..n\}$, Tu_i is item set of u_i, as defined in section 3.1.

- From the perspective of items, the relationship among users, tags and items is different, which is defined as $R_{P, UT}$. $R_{P, UT}= \{<p_k,UT_k>|p_k\in P, k=1..m\}$. UT_k is the user and tag set of item p_k. $UT_k= \{<u_i, t_j>| u_i\in U,t_j\in T$, and $E(u_i,t_j,p_k)=1\}$. Similarly, other two dimensional relationships $R_{P, U}$ and $R_{P, T}$ can be derived, which are defined as below:

$R_{P,\,U}$: The relationship between items and their user sets. Different from $R_{U,\,P}$ that describing each user's item set, $R_{P,\,U}$ is describing each item's user set. The traditional item-based collaborative filtering approach is based on this relationship. $R_{P,\,U}= \{<p_k, Up_k>| p_k \in P, Up_k \subseteq U, k=1..m\}$. Up_k is the user set of item p_k. $Up_k= \{u_i| u_i \in U, \exists t_j \in T, E(u_i,t_{j,} p_k) =1\}, p_k \in P, k=1..m$.

$R_{P,\,T}$: The relationship between items and their tag sets. $R_{P,\,T} = \{<p_k, Tp_k>| p_k \in P, Tp_k \subseteq T, k=1..m\}$ Tp_k is the tag set of item p_k. $Tp_k= \{t_i| t_j \in T, \exists u_i \in U, E(u_i,t_{j,} p_k) =1\}, p_k \in P, k=1..m$.

● From the perspective of tags, the relationship among users, tags and items is denoted as $R_{T,\,UP}$. Though it has not been used for the recommendation of items directly, we still give its definition as below for the sake of helping user get a whole view of the relationships among users, tags and items. $R_{T,\,UP}= \{<t_j,UP_j>| t_j \in T, j=1..l\}$. UP_j is the user and item set of tag t_j. $UP_j= \{<u_i, p_k>| u_i \in U,p_k \in P$, and $E(u_i,t_j,p_k)=1\}$. The other derived two-dimensional relationships $R_{T,\,U}$ and $R_{T,\,P}$ are defined as below:

$R_{T,\,U}$: The relationship between tags and their user sets. $R_{T,\,U}= \{<t_j, Ut_j>|t_j \in T, Ut_j \subseteq U, j=1..l\}$. Ut_j is the user set of tag t_j. $Ut_j=\{u_i| u_i \in U, \exists p_k \in P, E(u_i,t_{j,} p_k) =1\}, t_j \in T, j=1..l$.

$R_{T,\,P}$: The relationship between tags and their item sets. In this relationship, the tag collects all items that are being tagged with it by various users, which shows the result of this collaborative tagging work. $R_{T,\,P} = \{<t_j, Pt_j>| t_j \in T, Pt_j \subseteq P, j=1..l\}$ Pt_j is the item set of tag t_j. $Pt_j= \{p_k| p_k \in P, \exists u_i \in U, E(u_i,t_{j,} p_k) =1\}$.

These multiple relationships can be used to recommend personalized items, virtual friends, and tags to users. But for the scope of this paper, we will only focus on how to do item recommendations in the following sections.

3.3 Neighborhood Formation

Neighborhood formation is to generate a set of like-minded peers for a target user. Forming neighborhood for a target user $u_i \in U$ with standard "best-n-neighbors" technique involves computing the distances between u_i and all other users and selecting the top N 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 by using users' explicit rating data. However, explicit rating data is not always available. Unlike explicit ratings in which users are asked to supply their perceptions to items explicitly in a numeric scale, implicit ratings such as transaction histories, browsing histories, product mentions, etc., are also obtainable for most e-commerce sites and communities. For online communities with the tagging facility, binary implicit ratings can be obtained based on users' tagging information. If a user has tagged a product or item, the implicit rating to this item by this user is set to 1 otherwise 0.

For the implicit binary rating data, a simple but effective way to compute user similarity is to calculate the overlaps of two user's rated items. The higher the overlap, the more similar the two users are.

Based on the user profiles, two user's similarity is calculated. In Tso-Sutter's work, as the user was only profiled with the tag and item set, the similarity measure method of implicit rating behavior was used to form neighborhood. That is, the overlap of tags and items was used to measure the similarity [5]. However, it is not correct to measure the similarity of users' tagging behaviors as the same way as implicit rating behaviors. For example, for two users u_i and u_j with profiles

UF_i= ({globalization}, { The world is flat, The Long Tail}, {<globalization, The world is flat>, <globalization, The long Tail>}) and

UF_j= ({outsource, globalization}, {The world is flat, How Soccer Explains the World}, {<outsource, The world is flat>, <globalization, How Soccer Explains the World>}), the similarity measure should not only include the number of tags the users have used in common, the number of items the users have tagged in common, but also the number of using the same tag tagging the same item. If we just regard tagging behavior as implicit rating behavior, ignoring to measure the similarities of the relationships of tags and items, the wrong neighbors may be found. Only through calculating the similarity of tagging behaviors, the likely-minded users can be found.

Thus, the similarity measure of two users includes the following three parts:

(1) $UTsim(u_i,u_j)$: The similarity of users' tags, which is measured by the percentage of common tags used by the two users:

$$UTsim(u_i,u_j) = \frac{|Tu_i \cap Tu_j|}{\max_{u_k \in U}\{|Tu_k|\}} \tag{1}$$

As defined in section 3.1, Tu_i is the tag set of u_i, $Tu_i=\{t_j|t_j \in T, \exists p_k \in P, E(u_i,t_j, p_k) =1\}$

(2) $UPsim(u_i,u_j)$: the similarity of user's items, which is measured by the percentage of common items tagged by the two users:

$$UPsim(u_i,u_j) = \frac{|Pu_i \cap Pu_j|}{\max_{u_k \in U}\{|Pu_k|\}} \tag{2}$$

As defined in section 3.1, Pu_i is the item set of u_i, $Pu_i=\{p_k|p_k \in T, \exists t_j \in P, E(u_i,t_j, p_k) =1\}$

(3) $UTPsim(u_i,u_j)$: the similarity of the users' tag-item relationship, which is measured by the percentage of common relations shared by the two users:

$$UTPsim(u_i,u_j) = \frac{|TPu_i \cap TPu_j|}{\max_{u_k \in U}\{|TPu_k|\}} \tag{3}$$

As defined in section 3.1, TP_i is the relationship between u_i's tag and item set,$TP_i=\{<t_j, p_k>| t_j \in T, p_k \in P$, and $E(u_i,t_j,p_k)=1\}$

Thus, the similarity measure of two users is defined as below:

$$sim_u(u_i,u_j) = w_{UT} \cdot UTsim(u_i,u_j) + w_{UP} \cdot UPsim(u_i,u_j) + w_{UTP} \cdot UTPsim(u_i,u_j) \tag{4}$$

where $w_{UT} + w_{UP} + w_{UTP}=1$, w_{UT}, w_{UP} and w_{UTP} are the weighs to the three similarity measures, respectively. The weighs can be adjusted for different dataset. We can see the similarity measure of users is based on R_U that defined in section 3.2.

Similarly, the similarity between two items is based on R_P and is defined as formula (5) below:

$$sim_p(u_i, u_j) = w_{PU} \cdot PUsim(p_i, p_j) + w_{PT} \cdot PTsim(p_i, p_j) + w_{PUT} \cdot PUTsim(p_i, p_j) \quad (5)$$

where w_{PU}, w_{PT}, w_{PUT} =1, are the weights and their sum is 1, and PUsim(p_i, p_j), PTsim(p_i, p_j), PUTsim(p_i, p_j) are defined as follows:

(1) $PTsim(p_i, p_j)$: The similarity of two items based on the percentage of being put in the same tag, which is also computed based on the relationship $R_{U,T}$, but in the perspective of items.

$$PTsim(p_i, p_j) = \frac{|Tp_i \cap Tp_j|}{\max_{p_k \in P}\{|Tp_k|\}} \quad (6)$$

As defined in section 3.2, Tp_k is the tag set of item p_k, $Tp_k = \{t_i \mid t_j \in T, \exists u_i \in U, E(u_i, t_j, p_k) = 1\}$.

(2) $PUsim(p_i, p_j)$): the similarity of two items based on the percentage of being tagged by the same user, which is also calculated based on the relationship $R_{U,P,}$ and in the perspective of items.

$$PUsim(p_i, p_j) = \frac{|Up_i \cap Up_j|}{\max_{p_k \in P}\{|Up_k|\}} \quad (7)$$

As defined in section 3.2, Up_k is the user set of item p_k. $Up_k = \{u_i \mid u_i \in U, \exists t_j \in T, E(u_i, t_j, p_k) = 1\}$.

(3) $PUTsim(p_i, p_j)$: the similarity of the two items based on the percentage of common tag-item relationship, which is computed based on $R_{P,UT}$.

$$PUTsim(p_i, p_j) = \frac{|UP_i \cap UP_j|}{\max_{p_k \in P}\{|UP_k|\}} \quad (8)$$

As defined in section 3.2, UP_j is the user and item set of tag t_j. $UP_j = \{<u_i, p_k> \mid u_i \in U, p_k \in P, \text{ and } E(u_i, t_j, p_k) = 1\}$.

Though it's possible to calculate the similarity of two tags, it is not discussed in this paper.

3.4 Recommendation Generation

For a target user u_i, using the similarity measures discussed in section 3.3, we can generate the user's neighbourhood which contains users who have similar information needs or item preferences as u_i according to their tagging behaviour. We propose two methods to make item recommendations to the target user $u_{i,}$ namely, a user based approach and an item based approach, based on the neighbour users' item lists or the similarity of items, respectively.

Let $C(u_i)$ be the neighbourhood of u_i. For the user based approach, the candidate items for u_i are taken from the items tagged by the users in $C(u_i)$. For each candidate item p_k, based on the similarity between u_i and its neighbour users, and the neighbour users' implicit ratings to p_k that is denoted as $E(u_j, p_k)$, a prediction score denoted as $A^u(u_i, p_k)$ is calculated using Equation (9) given below. According to the prediction scores, the top N items will be recommended to $u_{i.}$.

For the item based approach, the prediction score is calculated by formula (10).

$$A^u(u_i,p_k) = \frac{\sum_{u_i\in C(u_i)} sim_u(u_i,u_j)\cdot E(u_j,p_k)}{|C(u_i)|} \tag{9}$$

$$A^p(u_i,p_k) = \sum_{p_j\in PP_{u_i}} sim_p(p_k,p_j) \tag{10}$$

4 Experiments

We have conducted experiments to evaluation the methods proposed in Section 3. The dataset for the experiments is obtained from Amazon.com. To avoid severe sparsity problem, we selected those users who tagged at least 5 items, tags that are used by at least 5 users, and items that are tagged at least 5 times. The final dataset comprises 3179 users, 8083 tags and 11942 books.

The whole dataset is split into a test dataset and a training dataset and the split percentage is 50% each. For each user in the testing dataset, a prediction score will be calculated for each item tagged by this user (i.e., the items which have implicit rating 1.). The top N items will be recommended to the user. The precision and recall are used to evaluate the accuracy of recommendations. If any item in the recommendation list has implicit rating 1 in the testing dataset, the item is counted as a hit.

To evaluate the effectiveness of the proposed tag based collaborative filtering approach, we compared the precision and recall of the recommended top 5 items of the proposed approach with the performance of the standard collaborative filtering (CF) approaches that only use the item information and also compared with Tso-Sutter's approach that extends the user rating matrix with the tag information. In fact, the proposed approach covers the two approaches when some of the similarity measure weights are set to zero. The comparison of precision and recall of user-based approaches is illustrated in Figure 1, while item-based comparison is shown in Figure 2.

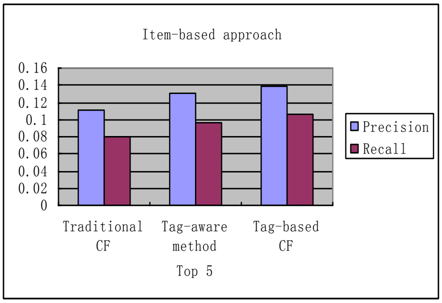

Fig. 1. Results of comparing the proposed Tag-based collaborative filtering employing user-based approach with user-based baseline model and the user-based Tag-ware approach proposed by Tso-Sutter

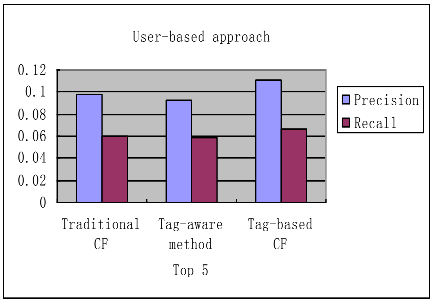

Fig. 2. Results of comparing the proposed Tag-based collaborative filtering employing item-based approach with item-based baseline model and the item-based Tag-ware approach proposed by Tso-Sutter

5 Discussion

The experiment results in Figure 1 and Figure 2 show that the precision and recall of the proposed approach are better than the traditional user- and item-based models and Tso-Sutter's approaches.

Though Tso-Sutter claimed that the tag information can only be useful to user and item fused collaborative filtering and it will be seen as noise for standard user- and item-based CF alone, our experiment results show that tag information can be used to improve the standard user-based and item-based collaborative filtering.

Besides, the experiment results also show that the traditional collaborative filtering recommendation based on the similarity of rating behavior doesn't work well to process the collaborative tagging information. The results suggest that it is more accurate and correct to profile user with tag, item and the relationship between tag and item than profiling user with extended implicit rating. Furthermore, the results also suggest that it is better to measure the similarity based on the similarity of tagging behaviour than just measuring it as implicit rating similarity.

6 Conclusion

This paper discusses how to recommend items to users based on collaborative tagging information. Instead of treating tagging behavior as just implicit rating behavior, the proposed tag based collaborative filtering approach uses the three dimensional relationship of tagging behavior to profile users and generate likely minded neighbors or similar items. The experiments show promising results of employing the tag based collaborative filtering approach to recommend personalized items. The experiment results also prove that the tag information can be used to improve the standard user-based and item-based collaborative filtering approaches.

References

1. Halpin, H., Robu, V., Shepherd, H.: The Complex Dynamics of Collaborative Tagging. In: The 16th international conference on World Wide Web, pp. 211–220. ACM, New York (2007)
2. Bao, S., Wu, X., Fei, B., Xue, S.Z., Yu, Y.: Optimizing Web Search Using Social Annotations. In: The 16th international conference on World Wide Web, pp. 501–510. ACM, New York (2007)
3. Marinho, L.B., Schmidt-Thieme, L.: Collaborative tag recommendations: Data Analysis, Machine Learning and Applications. In: The 31st Annual Conference of the Gesellschaft für Klassifikation, pp. 533–540. Springer, Heidelberg (2007)
4. Golder, S.A.: Usage patterns of collaborative tagging systems. Journal of Information Science 32(2), 198–208 (2006)
5. Tso-Sutter, K.H.L., Marinho, L.B., Schmidt-Thieme, L.: Tag-aware Recommender Systems by Fusion of Collaborative Filtering Algorithms. In: The 2008 ACM symposium on Applied computing, pp. 1995–1999. ACM, New York (2008)