Collaborative Filtering Recommender Systems Using Tag Information

Huizhi Liang, Yue Xu, Yuefeng Li, Richi Nayak
Faculty of Information Technology
Queensland University of Technology, Brisbane, Australia
oklianghuizi@gmail.com, {yue.xu, y2.li, r.nayak}@qut.edu.au

Abstract

Recommender Systems is one of the effective tools to deal with information overload issue. Similar with the explicit rating and other implicit rating behaviours such as purchase behaviour, click streams, and browsing history etc., the tagging information implies user's important personal interests and preferences information, which can be used to recommend personalized items to users. This paper is to explore how to utilize tagging information to do personalized recommendations. Based on the distinctive three dimensional relationships among users, tags and items, a new user profiling and similarity measure method is proposed. The experiments suggest that the proposed approach is better than the traditional collaborative filtering recommender systems using only rating data.

1. Introduction

Recommender systems can provide personalized contents, services and information items to potential consumers to decrease information retrieval time and support decision making process. Because user’s explicit rating is not always available, the implicit rating such as purchase history, downloading behaviour and click patterns etc. become another important information source for recommender systems.

With the development of web 2.0, collaborative tagging information becomes popular. Besides helping user organize 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 can be used to make recommendations.

Currently some researches are focusing on how to use collaborative tagging information to recommend personalized tags to users [2], but not much work has been done on utilizing tagging information to help users to find interested items easily and quickly.

In this paper, we will discuss how to recommend items to users based on tag information.

2. Related work

Collaborative filtering is a traditional and wildly used approach to recommend items to users based on the assumption that similar minded people may have similar taste or behaviors. In general, there are two kinds of collaborative filtering methods: user-based and item-based. Though there is a lot of work on the collaborative filtering recommender systems, to the best of our knowledge, only Tso-Sutter’s [3] work discussed about using the tag information to do item recommendation.

In Tso-Sutter’s work, the tag information was converted into two 2-dimensional relationships, user-tag and tag-item, and was used as a supplementary source to extend the rating data. Because it ignored the three dimensional relationship among users, items, and tags, the users’ tagging behavior was not accurately profiled, and thus the recommendation quality based on the extended data is still not satisfactory.

3. Tag-based Recommender systems
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. However, these approaches are used for describing two-dimensional relationships between users and items. Though Tso-Sutter’s approach takes the relationship between tags and items into consideration [3], it ignores the relationship between tags and items for each user. The user should be profiled not only by the tags and items, but also the relationship between the tags and items of the user.

To profile user’s tagging behavior 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 first:

\[ \begin{align*}
U &= \{u_1, u_2, \ldots, u_n\} : \text{Set of users in the collaborative tagging community.} \\
E &= \{p_1, p_2, \ldots, p_m\} : \text{Set of items that have been tagged by users.} \\
T &= \{t_1, t_2, \ldots, t_l\} : \text{Set of tags that have been used by users.} \\
E(u,t,p) &= \{0, 1\} : \text{a function that specifies whether user } u \text{ has used the tag } t \text{ to tag item } p. \\
\end{align*} \]

The user profile is defined as follows:

For a user \( u_i \), let \( T_{ui} \) be the tag set of \( u_i \), \( T_{ui} = \{t|t \in T \} \), \( P_{ui} \) be the item set of \( u_i \), \( P_{ui} = \{p|p \in P \} \), \( E(u_i, t, p) = \{0, 1\} \), \( E(u, t, p) = \{0, 1\} \), \( P_{ui} \subseteq P \), \( T_{ui} \subseteq T \). The user’s tagging behavior is defined as the user profile of user \( u_i \). The user profile or user model of all users is denoted as \( UF = \{U_i|i=1..n\} \).

3.2 Neighborhood Formation

Neighbourhood formation is to generate a set of like-minded peers for a target user. Based on user profiles, the similarity of users can be calculated through various proximity measures such as Pearson correlation and cosine similarity. In Tso-Sutter’s work, the overlap of tags shared by users was used to measure the similarity [3] just like the traditional collaborative filtering (CF) using the overlap of commonly rated items. Tso-Sutter’s method actually is the traditional CF with an extended dataset treating tags as additional items. The improvement is very limited, because the user’s neighborhood may be incorrectly formed if only treating users’ tagging as implicit rating and ignoring to measure the similarities of the relationships of tags and items. We propose to measure the similarity of two users from the following three aspects:

(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:

\[ UTsim(u_i, u_j) = \frac{|T_{ui} \cap T_{uj}|}{\max \{|T_{ui}|, |T_{uj}|\}} \]  

(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:

\[ UPSim(u_i, u_j) = \frac{|P_{ui} \cap P_{uj}|}{\max \{|P_{ui}|, |P_{uj}|\}} \]  

(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:

\[ UTPsim(u_i, u_j) = \frac{|TP_{ui} \cap TP_{uj}|}{\max \{|TP_{ui}|, |TP_{uj}|\}} \]  

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

\[ \text{Sim}(u_i, u_j) = w_{UT} \ast UTsim(u_i, u_j) + w_{UP} \ast UPSim(u_i, u_j) + w_{UTP} \ast UTPsim(u_i, u_j) \]  

Where \( w_{UT} \) and \( w_{UP} \) and \( w_{UTP} \) are the weights to the three similarity measures, respectively.

Similarly, the similarity between two items is defined
as formula (5).

\[
\text{Simp}(p_i, p_j) = w_{PU} \cdot \text{PUsim}(p_i, p_j) + w_{PT} \cdot \text{PTsim}(p_i, p_j)
\]

(5)

Where \(w_{PU}, w_{PT}, w_{PUT}\) are the weights. Their sum is 1, and \(\text{PUsim}(p_i, p_j), \text{PTsim}(p_i, p_j), \text{PUTsim}(p_i, p_j)\) are defined as follows:

1. \(\text{PTsim}(p_i, p_j)\): The similarity of two items based on the percentage of being put in the same tag.

\[
\text{PTsim}(p_i, p_j) = \frac{|T_{p_i} \cap T_{p_j}|}{\max\{|T_{p_i}|, |T_{p_j}|\}}
\]

(6)

Where \(T_{p_i}\) is the tag set of item \(p_i\). \(T_{p_i} = \{t_j | t_j \in T, E(p_i, t_j) = 1\}\).

2. \(\text{PUsim}(p_i, p_j)\): the similarity of two items based on the percentage of being tagged by the same user.

\[
\text{PUsim}(p_i, p_j) = \frac{|U_{p_i} \cap U_{p_j}|}{\max\{|U_{p_i}|, |U_{p_j}|\}}
\]

(7)

Where \(U_{p_k}\) is the user set of item \(p_k\). \(U_{p_k} = \{u_i | u_i \in U, \exists t_j \in T, E(u_i, t_j, p_k) = 1\}\).

3. \(\text{PUTsim}(p_i, p_j)\): the similarity of the two items based on the percentage of common tag-item relationship.

\[
\text{PUTsim}(p_i, p_j) = \frac{|U_{p_i} \cap U_{p_j}|}{\max\{|U_{p_i}||p_i \in P\}}
\]

(8)

Where \(U_{p_j}\) is the user and item set of tag \(t_j\). \(U_{p_j} = \{<u_i, p_k> | u_i \in U, p_k \in P, \text{ and } E(u_i, t_j, p_k) = 1\}\).

3.3 Recommendation Generation

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 \(p_k\) denoted as \(R(u_i, p_k)\), 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 \(u_i\).

\[
A^u(u_i, p_k) = \frac{\sum u_j \in C(u_i) \cdot R(u_j, p_k)}{|C(u_i)|}
\]

(9)

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

\[
A^p(u_i, p_k) = \sum \text{Simp}(p_k, p_j)
\]

(10)

4. Experiments

We have conducted experiments to evaluate the methods proposed in Section 3. The dataset for the experiments is obtained from Amazon.com. Because the items of the Amazon tagging community are mainly books, the book items are collected. 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. The top \(N\) items will be recommended to the user. The precision and recall are used to evaluate the accuracy of recommendations.

To evaluate the effectiveness of the proposed approach (Tag-based CF), we compared the precision and recall of the recommended top 5 items of the proposed approach with the performance of the standard collaborative filtering (Traditional CF) approaches that only use user ratings and also compared with Tso-Sutter’s approach (Tag-aware method) 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.

5. Discussion
The experimental results in Figure 1 and Figure 2 show that the precision and recall of the proposed approach are better than the traditional CF and Tso-Sutter’s approach for both user-based and item-based models.

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

![User-based approach](image)

Figure 1. Comparison of user based approaches

![Item-based approach](image)

Figure 2. Comparison of item-based approaches

Besides, the experimental results also show that the traditional collaborative filtering recommendation based on the similarity of rating behaviour doesn’t work well to process the collaborative tagging information. The results suggest that the recommendation accuracy is more improved by profiling users with tag, item and the relationship between tag and item than profiling users by extending implicit rating with tag information. Furthermore, the results also suggest that it is better to calculate the similarity based on the overall 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 behaviour as just implicit rating behaviour, the proposed tag based collaborative filtering approach uses the three dimensional relationship of tagging behaviour to profile users and generate most likely minded neighbours or similar items. The experiments show promising results of employing the tag based collaborative filtering approach to recommend personalized items. The experimental results also indicate that the tag information can be used to improve the standard user-based and item-based collaborative filtering approaches.

References:

