ABSTRACT

Precise recommendation of followers helps in improving the user experience and maintaining the prosperity of Twitter and microblog platforms. In this paper, we design a hybrid recommender system of microblog as a solution of KDD Cup 2012, track 1 task, which requires predicting users a user might follow in Tencent Microblog. We describe the background of the problem and present the algorithm consisting of keyword analysis, user taxonomy, (potential) interests extraction and item recommendation. Experimental result shows the high performance of our algorithm. Some possible improvements are discussed, which leads to further study.

General Terms

Algorithm, Machine Learning, Data Mining

Keywords

KDD Cup, keyword analysis, hybrid recommender system

1. INTRODUCTION

Online social networking services like Twitter have been tremendously popular, with a considerable speed of user growth. Thousands of new registrations are observed everyday in dominant platforms like Sina and Tencent Microblog since the introduction of microblog - Chinese twitter - in 2007. Celebrities and organizations also register microblog, which leads to diversity of topics and helps attract more potential users. However, flooded information can puzzle the users and even result in the loss of them. So reducing the risk of puzzlement and recommending attractive items - specific users selected for recommendation - are crucial for user experience improvement and prosperity maintenance, which present opportunities for novel machine learning and data mining approaches.

Recommender systems can be categorized into content-based algorithm [7], collaborative filtering [5], and influential ranking algorithm [12]. Unfortunately, all of them consider little of user profile’s fidelity, preference variance and interactions, causing difficulty of precise and stable recommendation. To overcome these weaknesses of single method, we construct a hybrid recommender system specified to Tencent Microblog, which generates ordered item list by mining the data of the platform [8].

The rest of the paper is organized as follows. Section 2 will discuss the background of the problem, and Section 3 will describe the design of the hybrid recommender system, including keyword analysis, user taxonomy, preference extraction, discovery of potential interests and generation of ordered recommendation. Section 4 will present the training process. Section 5 will show the experimental results and discuss some improvements, and the paper will be concluded in Section 6.

2. BACKGROUND

Observing the popularity of twitter services, Tencent, one of China’s leading Internet service portal, launched its microblog platform - Tencent Microblog - in 2010. It has attracted a lot of registered users (425 million registered accounts and 67 million daily active users in season 1, 2012 [11]) and became one of the dominant microblog platforms in China based on the large user group of its instant messaging service QQ (711.7 million [10] on Sep 30, 2011). Celebrities and organizations - items carefully categorized into hierarchies - are invited to register the platform, leading to a nice growth in the user group. Furthermore, Tencent’s microblog service is embedded in its other leading platforms like Shuoshuo (signature of user’s QQ account), Qzone (blog platform), Pengyou.com (SNS service) and Weixin (mobile messenger), hence user can write or comment a message directly on the website of Tencent Microblog or via the third-party port and related platforms.

While Tencent has the largest microblog user group, Sina Microblog takes a commanding lead with 56.5% of China’s microblog market based on active users and 86.6% based on browsing time over its competitors [6]. This fake prosperity of Tencent Microblog, which is far from the public perception, results from the existence of the fake users (explained in section 3.2), widely used spammer strategy [3] and the weird definition of active users. Tencent Microblog considers those who frequently write (retweet or comment) or read
microblog messages - no matter on the website or other associated platforms - as active users, while Twitter and Sina Microblog define them as those who login the platform everyday. User messages generated via from other related platforms confuse the recommender finding the real interests of users, which leads to the decrease of acceptance. Tencent Microblog users accept the recommendations in a low percentage (less than 9% according to our survey [8]), and the recommended item lists isn’t updated in time, which deviate from the users’ present preferences.

3. ALGORITHMS

As shown in the prior study, each of the existing algorithms has its unavoidable disabilities. This section introduces a hybrid recommender system to overcome them, including the preparations - keyword analysis and user taxonomy - and the main part of it.

3.1 Keyword Analysis

Mining synonyms in user’s keywords helps in finding their interests. However, applying association rule algorithm [2] to find them directly in the huge keyword set is unrealistic since that involves searching all possible combinations. So we parallel this process by adopting revised FDM(Fast Distributed Mining of association rules) [4], based on the downward-closure property of support which guarantees that the necessity and sufficient condition for a frequent itemset.

Distributed Mining of association rules) [4], based on the downward-closure property of support which guarantees that the necessity and sufficient condition for a frequent itemset.

The database DB(user-keyword set) is divided into n subsets DBi and these subsets are broadcasted to the remote sites RMi. Tj is the result generated in the jth iteration:

\[ T_j = \{ T_{i1}, T_{i2}, ..., T_{in} \} \]

where \( k_{ij} \) are the keywords of the jth keyword transaction in the jth iteration. Apriori algorithm is applied to generate the candidate transaction set \( C_i^j \) at RMi in the beginning of the jth iteration as follows:

\[ C_i^j = \text{Apriori\_gen}(T_{i}^{j-1}) \]

where

\[ C_i^j = \{ C_{i1}^j, C_{i2}^j, ..., C_{in}^j \} \]

\[ C_i^{j+1} = \{ k_{i1}^{j+1}, k_{i2}^{j+1}, ..., k_{im}^{j+1} \} \]

Let \( A = \{ A_1, A_2, ..., A_m \} \) be a set of mappings where

\[ A_j : \mathcal{K} \rightarrow \mathcal{W}_j \cup \{ 0 \} \]

\[ A_j(k) = \begin{cases} w_{ij}, & k_j = k_{ij} \in \mathcal{W}_j \\ 0, & k_i \notin \mathcal{W}_j \end{cases} \]

RMi computes the local support and confidence of \( C_i^j \) by

\[ \text{conf}_\text{local}(C_i^j) = \frac{\text{supp}_\text{local}(C_i^j)}{\text{supp}_\text{local}(C_i^{j-1})} \]

and eliminates those which fail to satisfy local minimums supp_local or conf_local. Then remote site sends the remaining candidate transactions \( C_i^j \) to the polling site \( \mathcal{P\mathcal{L}_K} \), where \( \mathcal{K} = \text{polling}(C_i^j) \) is a hash function.

\( \mathcal{P\mathcal{L}_K} \) gathers the candidate transactions \( C_i^j \) computes global support supp_global(\( C_i^j \)) and confidence conf_global(\( C_i^j \)) by sending request to remote sites for local values’ return:

\[ \text{supp}_\text{global}(C_i^j) = \text{average}(\min A_p(k_{i1}^j)) \]

\[ \text{conf}_\text{global}(C_i^j) = \frac{\text{supp}_\text{global}(C_i^j)}{\text{supp}_\text{global}(C_i^{j-1})} \]

Then \( \mathcal{P\mathcal{L}_K} \) filters out the candidates which fail to satisfy the constraint of supp_global and conf_global. Generally the local minimums coincide with the globals. For convenience we still denote the updated candidate sets as \( C_i^j \).

Then home site gathers \( C_i^j \) from the polling sites to generate the result of transactions(keyword classes) in the jth iteration:

\[ T_j^i = \bigcup C_i^j \]

where

\[ \text{supp}_\text{global}(T_j^i) >= \text{supp}_\text{global}, \]

\[ \text{conf}_\text{global}(T_j^i) >= \text{conf}_\text{global}. \]

The process is terminated if no new transaction is generated, before termination the home site broadcasts the transactions \( T_j^i \) to the remote site \( \mathcal{R\mathcal{M}_K} \) where \( \mathcal{R\mathcal{M}_K} \) is the original remote site of \( T_j^i = C_i^{K_m} \), and then starts next iteration.

The final result of keyword class is

\[ \text{keyword}_\text{class} = \{ \text{class}_1, \text{class}_2, ..., \text{class}_N \} \]

where

\[ \text{class}_i = \{ k_{i1}, k_{i2}, ..., k_{im} \} \]

is the set of synonyms.

The choice of minimum affects the precision and computational complexity tremendously. We sampled 1000 users’ keywords and found out that these users have their keyword weights average in 0.14, so we assign

\[ \text{supp}_\text{local} = \text{supp}_\text{global} = 0.2, \]

a little higher than the average weight. \( \text{conf}_\text{local}/\text{conf}_\text{global} \) is affected by supp_local/supp_global, in this case

\[ \text{conf}_\text{local} = \text{conf}_\text{global} = \frac{0.14}{0.2} = 0.7. \]
3.2 User Taxonomy

Most of the microblog platforms divide their users into 2 groups - active and inactive - to apply different types of strategies. However, some users don’t login Tencent Microblog directly while they have records of tweets generated from other related platforms, and these messages could hardly reflect their interests. Furthermore, they rarely interact with other users and have few favorites for the same reason. So we classify them as fake users(see Figure 1). In addition, we also consider the spammers as fake since they seldom use microblog even indirectly.

Due to the absence of login records, the activeness function $\text{act}(u_j)$ counts the number of tweets and interactions and computes $u_j$’s activeness by applying the thresholds $\text{min\_activeness}$ and $\text{min\_action}$:

$$\text{act}(u_j) = \text{tweet} \times \text{is\_fake}(u_j),$$

where

$$\text{is\_fake}(u_j) = 1 + \frac{\text{sgn(at + retweet + comment} - \text{min\_action})}{2}$$

and

$$\text{user\_class}(u_j) = \begin{cases} \text{active,} & \text{act}(u_j) \geq \text{min\_activeness} \\ \text{inactive,} & 0 < \text{act}(u_j) < \text{min\_activeness} \\ \text{fake,} & \text{act}(u_j) = 0 \end{cases}$$

We assign $\text{min\_activeness} = 100$ and $\text{min\_action} = 20$ since only 33.2% of the users have written more than 100 tweets(771599 in 2320895), and apply the algorithm to divide the user group into 3 classes.

An appropriate user taxonomy helps in improving the precision of recommendation. Users with similar favorites often accept similar items, hence dividing users into smaller groups by their interests can balance the precision and computational complexity. However, we haven’t done this due to the sparsity of successful recommendation records, which reflect the user’s interests directly.

3.3 Generating Recommendations

After keyword analysis and user taxonomy which are preparations of the recommendation, it comes the main part of our hybrid recommender system, consisting of item popularity ranking, (potential)interests discovery and the grading function, to generate recommended items and evaluate the possibility of acceptance or rejection. The system maps the users’ (potential)interests to their corresponding item categories and grades selected candidates in these categories with indicators of similarity and popularity. It also contains special algorithms with respect to fake users in order to reach a precise recommendation.

3.3.1 Item Popularity Ranking

An item is a specific user, which can be a famous person, an organization, or a group. Items are organized in different categories of professional domains by Tencent to form a hierarchy(see Figure 2). For example, an item, Dr. Kaifu LEE, is represented as science-and-technology.internet.mobile.1.

Due to the absence of login records, the activeness function $\text{act}(u_j)$ counts the number of tweets and interactions and computes $u_j$’s activeness by applying the thresholds $\text{min\_activeness}$ and $\text{min\_action}$:

$$\text{act}(u_j) = \text{tweet} \times \text{is\_fake}(u_j),$$

where

$$\text{is\_fake}(u_j) = 1 + \frac{\text{sgn(at + retweet + comment} - \text{min\_action})}{2}$$

and

$$\text{user\_class}(u_j) = \begin{cases} \text{active,} & \text{act}(u_j) \geq \text{min\_activeness} \\ \text{inactive,} & 0 < \text{act}(u_j) < \text{min\_activeness} \\ \text{fake,} & \text{act}(u_j) = 0 \end{cases}$$

We assign $\text{min\_activeness} = 100$ and $\text{min\_action} = 20$ since only 33.2% of the users have written more than 100 tweets(771599 in 2320895), and apply the algorithm to divide the user group into 3 classes.

An appropriate user taxonomy helps in improving the precision of recommendation. Users with similar favorites often accept similar items, hence dividing users into smaller groups by their interests can balance the precision and computational complexity. However, we haven’t done this due to the sparsity of successful recommendation records, which reflect the user’s interests directly.

3.3 Generating Recommendations

After keyword analysis and user taxonomy which are preparations of the recommendation, it comes the main part of our hybrid recommender system, consisting of item popularity ranking, (potential)interests discovery and the grading function, to generate recommended items and evaluate the possibility of acceptance or rejection. The system maps the users’ (potential)interests to their corresponding item categories and grades selected candidates in these categories with indicators of similarity and popularity. It also contains special algorithms with respect to fake users in order to reach a precise recommendation.

3.3.1 Item Popularity Ranking

An item is a specific user, which can be a famous person, an organization, or a group. Items are organized in different categories of professional domains by Tencent to form a hierarchy(see Figure 2). For example, an item, Dr. Kaifu LEE, is represented as science-and-technology.internet.mobile.1.

Due to the absence of login records, the activeness function $\text{act}(u_j)$ counts the number of tweets and interactions and computes $u_j$’s activeness by applying the thresholds $\text{min\_activeness}$ and $\text{min\_action}$:

$$\text{act}(u_j) = \text{tweet} \times \text{is\_fake}(u_j),$$

where

$$\text{is\_fake}(u_j) = 1 + \frac{\text{sgn(at + retweet + comment} - \text{min\_action})}{2}$$

and

$$\text{user\_class}(u_j) = \begin{cases} \text{active,} & \text{act}(u_j) \geq \text{min\_activeness} \\ \text{inactive,} & 0 < \text{act}(u_j) < \text{min\_activeness} \\ \text{fake,} & \text{act}(u_j) = 0 \end{cases}$$

We assign $\text{min\_activeness} = 100$ and $\text{min\_action} = 20$ since only 33.2% of the users have written more than 100 tweets(771599 in 2320895), and apply the algorithm to divide the user group into 3 classes.

An appropriate user taxonomy helps in improving the precision of recommendation. Users with similar favorites often accept similar items, hence dividing users into smaller groups by their interests can balance the precision and computational complexity. However, we haven’t done this due to the sparsity of successful recommendation records, which reflect the user’s interests directly.

3.3 Generating Recommendations

After keyword analysis and user taxonomy which are preparations of the recommendation, it comes the main part of our hybrid recommender system, consisting of item popularity ranking, (potential)interests discovery and the grading function, to generate recommended items and evaluate the possibility of acceptance or rejection. The system maps the users’ (potential)interests to their corresponding item categories and grades selected candidates in these categories with indicators of similarity and popularity. It also contains special algorithms with respect to fake users in order to reach a precise recommendation.

3.3.1 Item Popularity Ranking

An item is a specific user, which can be a famous person, an organization, or a group. Items are organized in different categories of professional domains by Tencent to form a hierarchy(see Figure 2). For example, an item, Dr. Kaifu LEE, is represented as science-and-technology.internet.mobile.1.

Due to the absence of login records, the activeness function $\text{act}(u_j)$ counts the number of tweets and interactions and computes $u_j$’s activeness by applying the thresholds $\text{min\_activeness}$ and $\text{min\_action}$:

$$\text{act}(u_j) = \text{tweet} \times \text{is\_fake}(u_j),$$

where

$$\text{is\_fake}(u_j) = 1 + \frac{\text{sgn(at + retweet + comment} - \text{min\_action})}{2}$$

and

$$\text{user\_class}(u_j) = \begin{cases} \text{active,} & \text{act}(u_j) \geq \text{min\_activeness} \\ \text{inactive,} & 0 < \text{act}(u_j) < \text{min\_activeness} \\ \text{fake,} & \text{act}(u_j) = 0 \end{cases}$$

We assign $\text{min\_activeness} = 100$ and $\text{min\_action} = 20$ since only 33.2% of the users have written more than 100 tweets(771599 in 2320895), and apply the algorithm to divide the user group into 3 classes.

An appropriate user taxonomy helps in improving the precision of recommendation. Users with similar favorites often accept similar items, hence dividing users into smaller groups by their interests can balance the precision and computational complexity. However, we haven’t done this due to the sparsity of successful recommendation records, which reflect the user’s interests directly.
After keyword analysis of individuals the target categories hence the candidate items are the items in each category is defined to construct the keyword class of a given category. A mapping generated by the keyword analysis. A mapping computes the keyword class of a given user satisfies relation to their followees. We build up the social network of a user and search for its potential interests in its followees followee(υj) = {υk} and even in its followees’ followees(see Figure 3).

Let depth be the maximal levels amount of the searching process, in fact depth ≤3 is enough for the process to mine a user’s potential interests. The related users of υj is

\[
\text{related\_users(υj)} = \text{search\_followee(υj, depth)}
\]

where search\_followee(υj, depth) returns the followees and indirect followees of υj with max level depth in the social network. Then for every υk in related\_users(υj) we compute the keyword classes as mentioned above and merge them into the set

\[
\text{potential\_key(υj)} = \{\text{class}_1, \text{class}_2, ..., \text{class}_j\} = \bigcup_{υk \in \text{related\_users(υj)}} \text{key\_class(υk)}
\]

where the \text{i}th keyword class \text{class}_j has the weight

\[
\tilde{W}_{ji} = \sum_{υk \in \text{related\_users(υj)}} W_{kl} \text{fami}(υj, υk).
\]

\text{fami}(υj, υk) computes the familiarity of υj and υk by adopting indicators of interactions(at(0), retweet and comment) which could only happen in linked users:

\[
f(υj, υk) = \omega_1 f(at) + \omega_2 f(retweet) + \omega_3 f(comment),
\]

where \omega_i satisfying \sum_{i=1}^3 \omega_i = 1, \omega_i ≥ 0 is obtained in training process and \(f(x)\) is a sigmoid function

\[
f(x) = \frac{2}{1 + e^{-x}} - 1.
\]

Finally we merge key\_class(υj) and potential\_key(υj) into the set

\[
\text{interests}(υj) = \{\text{class}_1, \text{class}_2, ..., \text{class}_j\}
\]

with weight of \text{class}_j

\[
W_{ji} = \begin{cases} W_{j1}, & \text{class}_j \in \text{key\_class(υj)} \\ W_{j2}, & \text{class}_j \in \text{potential\_key(υj)} \\ \frac{1}{2}(W_{j1} + W_{j2}), & \text{class}_j in both sets \end{cases}
\]

(here \(υ_j\) can be substituted by \(υ_k\)), where

\[
\text{weight}_i = \begin{cases} W_{ji}, & \text{class}_j = \text{class}_i \in \text{key\_class(υ_j)} \\ 0, & \text{class}_j \notin \text{key\_class(υ_j)} \end{cases}
\]

The similarity between \(υ_k\) and \(υ_j\) is the normalized Euclid distance of these 2 vectors:

\[
\text{sim}(υ_j, υ_k) = \frac{\|\text{class\_weight(υ_j)} - \text{class\_weight(υ_k)}\|}{\text{size}}
\]

where \(\text{size}\) is the number of k in both sets.

### 3.3.2 Mining Interests from Keywords

Users are inclined to accept items of their interests. Active users have more keywords which reflect their favorites, and we map these interests to the hierarchy to obtain candidate items.

Consider the keyword classes set

\[
\text{keyword\_class} = \{\text{class}_1, \text{class}_2, ..., \text{class}_N\}
\]

generated by the keyword analysis. A mapping

\[
\text{KH} : H \rightarrow \text{power(\text{keyword\_class})}
\]

is defined to construct the keyword class of a given category \(h_κ\) (see section 3.3.4 for details). Suppose a given user \(υ_j\) (or a given item \(υ_k\)) has keywords \(K_j = \{k_1, k_2, ..., k_n\}\) with weights \(w_1, w_2, ..., w_n\) which satisfy \(\sum_{i=1}^n w_i = 1\). A function

\[
\text{key\_class}(υ_j) = \{\text{class}_j, \text{class}_j, ..., \text{class}_j\}
\]

computes the keyword class of a given user \(υ_j\) (or item \(υ_k\)) where \(\text{class}_j\) satisfies

\[
K_j \cap \text{class}_j \neq \emptyset,
\]

and the corresponding weight of \(\text{class}_j\) is

\[
\tilde{W}_{ji} = \sum_{k_i \in K_j \cap \text{class}_j} w_i.
\]

After keyword analysis of individuals the target categories \(\{h_κ\}\) is generated where \(h_κ\) satisfies

\[
\text{KH}(h_κ) \cap \text{key\_class(υ_j)} \neq \emptyset,
\]

hence the candidate items are the items in each \(h_k\) (suppose \(υ_k\) included). A vector function is defined on \(U \cup I\) as follows:

\[
\text{class\_weight}(υ_j) = (\text{weight}_1, \text{weight}_2, ..., \text{weight}_N)
\]
where
\[ class_{ji} = class_{ji1} \in \text{key_class}(u_j), \]
\[ class_{ji} = class_{ji2} \in \text{potential_key}(u_j). \]
Correspondingly the target category \( \{ h_k \} \) satisfies
\[ KH(h_k) \cap \text{interests}(u_j) \neq \emptyset, \]
and the candidate items are in each \( h_k \) as mentioned before (suppose \( i_k \) included). We modify the vector function \( \text{class_weight}(u_j) \) which is defined in section 3.3.2 by using \( \text{interests}(u_j) \) to substitute \( \text{key_class}(u_j) \), i.e. \( W_{ji} \) instead of \( W_{ji} \), and compute the similarity of \( u_j \) and \( i_k \) as before. In this way, we get the similarity between items and inactive users. The algorithm can also be applied to the recommendation for the active users with smaller value of depth.

### 3.3.4 Grading Function

In studies above we presented the extraction of a given user \( u_j \)'s (potential) interests and the generation of the recommended candidates. The grading function \( grade(u_j, i_k) \) computes the possibility of acceptance (positive grade) or rejection (negative grade) with indicators of \( i_k \)'s popularity and \( \text{sim}(u_j, i_k) \) computed as above (see Figure 4). Then we pick out the first \( k \) candidates and sort them in descending order to generate final recommendation, where in our case \( k = 3 \).

Let \( I = \{ i_1, i_2, \ldots, i_n \} \) and \( H = \{ h_1, h_2, \ldots, h_n \} \) be the item set and the category (hierarchy) set as previously defined and suppose we have extracted keyword classes of each user and item. We specify the definition of \( KH \) (see section 3.3.2) as follows:

\[ KH : H \to \text{power(keyword\_class)}, \]
\[ KH(h_k) = \{ class_{h_k1}, class_{h_k2}, \ldots, class_{h_kn_k} \} \]
\( KH(h_k) \) computes the keyword classes of a given category \( h_k \) with corresponding weight of \( class_{h_k} \)
\[ W_{hp} = \text{average}(W_{ji}), \]
\[ i_j \in h_k, \text{key_class}(i_j)(l_j) = class_{ji} = class_{h_k}. \]
Here \( \text{power(keyword\_class)} \) is the power set of \( \text{keyword\_class} \).

We revise the definition of \( \text{class_weight}() \) by extending its domain to \( H \times H \):

\[ \text{class_weight}(h_k) = (weight_1, weight_2, \ldots, weight_n), \]
\[ weight_i = \left\{ \begin{array}{ll}
W_{kp}, & \text{class}_i = class_{hp} \in KH(h_k) \\
0, & \text{class}_i \notin KH(h_k)
\end{array} \right. \]

As previously mentioned we don't compute these vectors directly. The function \( fond() \) is defined on \( U \times H \) by
\[ fond(u_j, h_k) = g(\text{class_weight}(u_j) \cdot \text{class_weight}(h_k), 100), \]
\[ g(x, y) = \frac{2(1 + e^{-y})}{(1 - e^{-y})(1 + e^{-x})} - \frac{1 + e^{-y}}{1 - e^{-y}} \]
to compute the ratio of \( u_j \)'s fondness for category \( h_k \).

Finally the grading function of active/inactive users is
\[ grade(u_j, i_k) = 2fond(u_j, h_k)(\alpha_1 hot_k + \alpha_2 \text{sim}(u_j, i_k)) - 1, \]
where \( hot_k, sim(u_j, i_k) \) are indicators as mentioned above and \( \alpha_1, i = 1, 2 \) satisfying \( \alpha_1 + \alpha_2 = 1, \alpha_1 \geq 0 \) are obtained in the training process. Valued in \([-1, 1]\), the grading shows the possibility of acceptance (positive grade) or rejection (negative grade). Considering the variance of user preferences in a certain period, a revised grading function is defined as
\[ \text{revised grade}(u_j, i_k) = \frac{1}{\lambda} \text{time}(u_j, h_k) \text{grade}(u_j, i_k), \]
where \( \text{time}(u_j, h_k) = 1 + (\lambda - 1)e^t, t \in (-\infty, 0] \).

Let \( result(u_j, i_k) \) be a record of recommendation in the training set where
\[ result(u_j, i_k) = \begin{cases} 
1, & \text{recommendation accepted} \\
-1, & \text{other} \end{cases} \]
Consider the training error in the \( n \)th epoch
\[ error(n) = result(u_j, i_k) - \text{grade}(u_j, i_k). \]
The gradients (partial derivatives) of \( error(n) \) are
\[ \frac{\partial}{\partial \alpha_1} \text{error}(n) = 2fond(u_j, h_k)(\text{hot}_k - \text{sim}(u_j, i_k)), \]
\[ \frac{\partial}{\partial \omega_1} \text{error}(n) = 2\alpha_2 \text{fond}(u_j, h_k) \sum_{l=1}^{m_1} \frac{\partial W_{jl}}{\partial \omega_1}, \]
\[ \frac{\partial}{\partial \omega_2} \text{error}(n) = 2\alpha_2 \text{fond}(u_j, h_k) \sum_{l=1}^{m_2} \frac{\partial W_{jl}}{\partial \omega_2}. \]
Figure 4: Recommendation process. Keywords analysis extracts user’s interests, $KH$ mapping match these interests to corresponding item categories thus obtains the set of candidates, and the grading function assigns grades of the recommendation by computing the similarity and popularity.

where

$$\sigma = sgn(class\_weight(u_j) - class\_weight(i_k)),$$

$$W_{jl} = W_{jl1} + W_{jl2}.$$

$W_{jl1}$ is the weight of $u_j$’s own keyword class and $W_{jl2}$ is computed by searching related users (section 3.3.3). By rounding $\partial W_{jl2}$, $fami(u_j, u_k)$:

$$\frac{\partial W_{jl1}}{\partial \omega_2} = \frac{\partial W_{jl2}}{\partial \omega_2} + \sum_{u_k \in related\_users(u_j), class_k = class_{l2}} W_{klk}(f(retweet) - f(comment)).$$

After computation of the gradients we update $\alpha_1$ and $\omega_1(i = 1, 2)$ by applying momentum factor $\beta \in [0, 1]$. The parameter $\alpha_1$ in the $n^{th}$ epoch $\alpha_1(n)$ satisfies

$$\alpha_1(n + 1) = \alpha_1(n) + (1 - \beta) \frac{\partial error(n)}{\partial \alpha_1} + \beta \Delta \alpha_1(n),$$

where

$$\Delta \alpha_1(n) = \alpha_1(n) - \alpha_1(n - 1).$$

$\omega_1, i = 1, 2$ are updated similarly. The training process is terminated in the $n^{th}$ epoch if no new training data $result(u_j, i_k)$ or $error(n) \leq performance$, where $performance = 0.01$ controls the training process.

Furthermore, if we apply $\text{revised}\_grade()$ instead of $\text{grade}()$ in the grading process, the gradients should be revised as

$$\frac{\partial error(n)}{\partial \omega_1} = \frac{2}{\lambda} \sigma_{\text{time}}(u_j, h_k) fond(u_j, h_k) (hot_k - \text{sim}(u_j, i_k)),$$

$$\frac{\partial error(n)}{\partial \omega_2} = \frac{2}{\lambda} \sigma_{\text{time}}(u_j, h_k) fond(u_j, h_k) \sum_{l=1}^{m} \frac{\partial W_{jl}}{\partial \omega_2}.$$

5. EXPERIMENT AND IMPROVEMENT

This section discusses the training result and evaluation metrics we adopted to test the system’s performance. Some of other approaches are also introduced, which might enhance the performance of the recommender system.

5.1 Training Result

In our experiment we sampled 5,938 users’ recommendation records from the dataset [8] stochastically and divided them into 2 subsets for training and testing. We omitted the update of $\omega_1$ to reduce the computational complexity and assigned $\omega_1 = \frac{1}{2}$, which means at $(0)$, retweet and comment occupy the same proportion when computing $fami(u_j, u_k)$. We trained each user’s patterns and computed the average parameters. Table 1 presents the results of the training process.

| user class | user followee interaction keyword | $\alpha_1$ |
|------------|---------------------------------|-----------|
| active     | 3919                            | 46        | 87       | 10 | 0.33 |
| inactive   | 1194                            | 27        | 42       | 8  | 0.18 |
| fake       | 825                             | 18        | 2        | 5  | /   |

Table 1: Training Sets and Optimal Parameters. Fake users’ grading function has no parameters to update so we omit the training process of it.

The result shows an evident discrepancy of $\alpha_1$, which reflects the inclination of accepting popular items. Inactive users prefer items with similar interests while active users prefer items with high popularity.

5.2 Prediction and Precision Evaluation

We computed $\text{grade}(u_j, i_k)$ of all $\text{result}(u_j, i_k)$ in testing subset and generated ordered item list of $u_j$ (see section 3.3.4) to test the trained system. The evaluation metric is the average precision [14] which KDD Cup’s organizers adopted:

$$AP@3(u_j) = \sum_{i=1}^{3} p(i) \Delta r(i),$$

where $p(i)$ is the precision of the $i^{th}$ recommended item and $\Delta r(i)$ is the change in the recall from $i - 1$ to $i$. Table 2 presents the MAP@3 (mean value of $AP@3(u_j)$) results and Table 3 presents the recommended item lists and the average precision of some users. The precision of fake users’ prediction is much lower than others’ in our experiment due to the difficulty of their interests’ extractions. Adjusting $\text{min\_action}$ or recommending their linkers on other related platforms like QQ might help improve the results.
Table 2: Prediction Evaluation. Mining potential interests from inactive users’ followees improves the performance of recommendation. Fake users’ result is not good as the others.

| user class | item accepted item | AP@3(u_j) |
|------------|--------------------|-----------|
| active     | 2071402            | 0.83      |
|            | 1606902            | 1760350   |
|            | 1774452            |           |
| inactive   | 942226             | 1.00      |
|            | 1606902            | 1606609   |
|            | 1774452            |           |
| fake       | 193889             | 0.33      |
|            | 1760642            | 1774862   |
|            | 1774862            |           |

Table 3: Examples of Prediction. User 2071402 accepts the 1st and 3rd items, then $AP@3 = \left( \frac{1}{3} + \frac{2}{3} \right) / 2 = \frac{2}{3}$; User 942226 only accepts the 1st item, then $AP@3 = \frac{1}{3} = 1$; User 193889 only accepts the 3rd item, then $AP@3 = \frac{1}{3}$.

5.3 Improvements of the System

There are approaches to enhance the performance and overcome the limitations of our system. Recommendation based on demographic methods can help in enhancing the percentage of acceptance. Users with similar demographic information may have interest’s coincidences thus accept similar items. Refined keyword analysis and user taxonomy can improve the recommendation. Users who follow items in the same category or interact with users who have explicit preferences can be grouped in identical user class. They share synonyms in their keywords and accept similar items in a high possibility based on the similarity of preferences. Adaptation to the frequently updated microblog platform’s database can get user’s present interests. Fortunately user’s interests and behaviors are stable in a short period, so the system only needs retrain stochastically and gradually, which is fast and accurate.

6. CONCLUSION AND FUTURE WORK

We present a hybrid recommender system for microblog to solve Track 1 task, KDD Cup 2012. The system analyzes the synonyms of keywords and behaviors of different users, extracts their (potential)interests, finds the target categories, grades the candidate items in those categories with indicators of popularity and similarity, and finally generates ordered item lists respect to each user. Experimental result shows high performance of our algorithm. The initialization of grading function’s parameters needs improvement, since good choices of them accelerate the process and avoid local minimums. Dynamic algorithms which reduce the risk of inaccuracy by searching the best algorithm through competition also deserves further study.

7. ACKNOWLEDGMENTS

We would like to thank the organizers of KDD Cup 2012 for organizing such a challenging and exciting competition. We would also like to thank Jiuya Wang for helpful discussions and Dachao Li from High-Performance Computing Platform of Sun Yat-Sen University for distributed computing supports.

8. REFERENCES

[1] K. C. 2012. Predict which users (or information sources) one user might follow in tencent weibo. http://www.kddcup2012.org/c/kddcup2012-track1, 2012.
[2] R. Agrawal, T. Imieliński, and A. Swami. Mining association rules between sets of items in large databases. In Proceedings of the 1993 ACM SIGMOD international conference on Management of data, SIGMOD ’93, 1993.
[3] Baidu. Zombie fans on weibo. http://baike.baidu.com/view/4047998.htm, 2010.
[4] D. Cheung, J. Han, V. Ng, A. Fu, and Y. Fu. A fast distributed algorithm for mining association rules. In Parallel and Distributed Information Systems, 1996., Fourth International Conference on, dec 1996.
[5] J. A. Konstan, B. N. Miller, D. Maltz, J. L. Herlocker, L. R. Gordon, and J. Riedl. GroupLens: applying collaborative filtering to usenet news. Commun. ACM, 1997.
[6] Kyle. Sina commands 56% of china’s microblog market. http://www.resonancechina.com/2011/03/30/sina-commands-56-of-chinas-microblog-market/, March 2011.
[7] M. McPherson, L. Smith-Lovin, and J. M. Cook. Birds of a feather: Homophily in social networks. Annual Review of Sociology, 2001.
[8] Y. Niu, Y. Wang, G. Sun, A. Y. B. Dalessandro, C. Perlich, and B. Hamner. The Tencent Dataset and KDD-Cup’12, KDD-Cup Workshop, 2012.
[9] P. Tan, M. Steinbach, and V. Kumar. Introduction to Data Mining, chapter 6. Association Analysis: Basic Concepts and Algorithms. Addison-Wesley, 2005.
[10] Tencent. About tencent. http://www.tencent.com/en-us/at/abouttencent.shtml, 2012.
[11] Tencent. Tencent announces 2012 first quarter results. http://www.tencent.com/en-us/content/ir/news/2012/attachments/20120516.pdf, May 2012.
[12] Z. Wang, Y. Tan, and M. Zhang. Graph-based recommendation on social networks. In Web Conference (APWEB), 2010 12th International Asia-Pacific, 2010.
[13] Wikipedia. Demography. http://en.wikipedia.org/wiki/Demography.
[14] M. Zhu. Recall, precision and average precision. Working Paper 2004-09, Department of Actuarial Science, University of Waterloo, 2004.