A Hybrid Friend Recommendation Technique Using SVM Based on the Users' Attributes

Liu Yang, Deborah M. Ferreira, Jianya Zheng, Li Weigang

Department of Computer Science – University of Brasilia
Brasilia, Brazil

gilbertoliu514@msn.com, deborahmdf@gmail.com, darcy_zheng@hotmail.com, weigang@cic.unb.br

Abstract. *Online social networks have attracted millions of users to integrate their daily life into these social media. As an important e-activity, the Friend Recommendation System (FRS) has been designed to help the users exploring new friends with common interests. However, most existing FRS are using simple methods, such as mutual friends, location-based information etc. This paper proposes a hybrid technique utilizing Support Vector Machine (SVM), recommending people in social networks based on users' attributes. With the case study from Tencent Weibo, the proposed method has improved the accuracy of recommendation comparing with two classic algorithms, Naive Bayes and Random Forests. Furthermore, considering nine attributes in FRS, we identified that the acceptance of followers is the most important factor to influence the users' decision.*

1. Introduction

Online social networks (OSNs) have grown by leaps and bounds with the emergence of web2.0 over past years. The microblog, as one of the most popular form of OSN, has expanded rapidly due to their convenient form when adapting to the modern mobile terminal (for example, cell phone, tablet, etc.). Furthermore, E-commerce and business review sites pay increasingly attention on the combination of the social network information to increase their sales and reliable comments [Swamynathan et al 2008]. They believe that the content of social network is very valuable in the future.

FRS have been widely used in the social networks, such as Facebook and Twitter. But the existing friend recommendation systems are too simple to provide quality recommendations. Most of them only employ "mutual friends" or location-based information, etc[Ma et al 2011].

Until now, the application of support vector machine (SVM) in online social networks is at the early stage, few work focuses on the friend recommendation. Ting [Ting et al 2012] proposed the architecture of a social recommendation system based on the data from microblogs. The similarity of the discovered features of users and products will then be calculated as the essence of the recommendation engine. Tian[Tian and Zheng 2012] focus on using Sina Weibo, the most popular Chinese microblogging platform, for the task of user character analysis. They defined four categories features by analyzing microblogs, and show how to collect labeled corpus as training data. Using the corpus and via SVM, they build a character classifier, which is able to determine extraversion or introversion in a microblog. The experimental

evaluation shows that their method can identify user's character accurately and efficiently.

In this study, we propose a hybrid approach based on the SVM to perform the friend recommendation in OSNs. Nine attributes are abstracted from the users' profile to study the influence of the user's acceptance. With the case study in Tencent Weibo, the proposed method can improve the accuracy of recommendation comparing with two classic algorithms, Naive Bayes and Random Forests. Furthermore, within the nine attributes we studied in this work, the acceptance of followers is the most important one in influencing users' decision.

2. Attributes and Model

In machine learning and statistics, classification is the problem of identifying to which set a group of instances belongs, based on some specific rules. In this study, the friend recommendation is considered as a binary classification problem. We represent the user's acceptance of recommendation by 1 and rejection by -1. All instances in the training set are separated into two categories. We use the experimental data offered by the Tencent Weibo, the biggest microblog social network in China. The dataset size is 3.1GB with 73 million instances. There are 2.3 million users and 6 thousands items in the dataset. The item is a special user which can be a person, an organization, or a group, that was selected and recommended to other users. Typically, items are celebrities, famous organizations, or some well-known groups. For example, the singer star Lady Gaga is represented as an item, she is recommended to the users. If the user accepts the recommendation, it will be expressed as 1 in training set.

2.1. Attributes Selection

The profile of microblog users has several attributes which can influence on their social relationship. For example, the number of fans which the item owns can directly reflect its popularity in social network. Besides, the common followers that your friends have is probably a person interesting to you. Thus, we choose nine attributes as features to train the classifier. There are Gender, Age, Activity, Relative network, Keyword similarity, Item category, followee acceptance, follower acceptance.

2.2. Attributes' Model

There are many attributes can be retrieved directly such as age, gender, number of fans, category. Other attributes must go through a particular calculation process before we get it. The follower acceptance is calculated by percentage of user acceptance based on the history record of user's acceptance. For example, if the system recommends 100 items to a user and the user accepts 30 of them, therefore, the user's follower acceptance will be 30%. Equally, the followee acceptance was calculated by percentage of items that were accepted. All of this data offered by the Tencent Weibo.

There are three attributes that need to design a model for each one. They are: relative network, keyword similarity and activity.

A: Relative Network

In microblog, the relationships between users can be described by Follow Model [Sandes et al 2012]. First, we define the relevant conceptions of 'follower' and 'followee'. If A is B's fan, then A is B's follower and B is A's followee. E represents the connection between different nodes in the social relationship network. Follow Model proposed three functions to express the relationships in social network.

$$\begin{cases} f_{out}(A) = \{B \mid (A,B) \in E\} & \text{The person whom } A \text{ followed.} \\ f_{in}(B) = \{A \mid (A,B) \in E\} & \text{The person who followed by } A. \end{cases}$$

We can use these two functions to combine the relationship between users to build the recommendation relationship.

1) $f_{out}(A) \cap f_{out}(B)$: common followees of two users.
2) $f_{in}(A) \cap f_{in}(B)$: common followers of two users.
3) $f_{out}(A) \cap f_{in}(B)$: intersection of the followees of A and followers of user B.

Among these relationships, relationship 3 is the most suitable for data provided by the experiment, we describe it in Figure 1. The number of users they owns is one of the attributes.

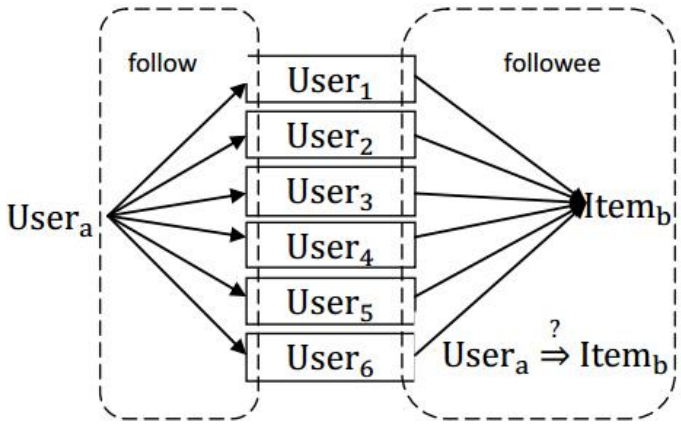

Figure 1. Relationship $f_{out}(A) \cap f_{in}(B)$

B: Keyword Similarity

The keywords and their weights are given in the dataset. The keyword represents the user's characteristic and interests. The similarity of keyword can represent the similarity between different users[Li et al 2008]. We calculate the keyword similarity between user and item. The equation is as follows:

$$Sim(A,B) = \frac{\sum_{k \in K} w_{Ak} \times w_{Bk}}{\sum_{k \in K} (w_{Ak})^2 \times \sum_{k \in K} (w_{Bk})^2}$$

Where w_{AK} represent the weight of keyword of user A, with more than six hundred thousand keywords, we found that only less than 1% pairs of users have the same keywords. So the keyword similarity cannot be used as an attribute directly. To resolve the data sparsity problem, only the similarity between items is calculated. We figure out the items that the user already followed and calculate the keyword similarity successively between pairs of item. Finally we summed all of the value of the pair.

C: Activity

User activity can be measured by their number of microblog, number of retweet and number of comment. The simplest measurement is summing all these actions, but this approach is flawed. Because it is not precise and can be easily manipulated, for example, one user that retweets one thousand times, but posts and comments less than 10 times, this is called "zombie fans". They are produced by spam programs and make no sense for our research. There are two approaches to treat this problem. One is based on expert's experience and this approach is relied on the advice of expert. As our experiments is not able to get expert advice. Another approach "objective weighting method" [Shechtman 2013]is adopted to calculate the activity of users.

$$CV = \frac{\sigma}{\mu} = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}}{(x_1 + x_2 + ... + x_n)/N}$$

Where σ represents the standard deviation(number of microblog, number of retweet or number of comment), μ represents the mean value(number of microblog, number of retweet or number of comment). This equation is usually used as a measurement when parameters are not in equal situations.

3. Modeling the Friend Recommendation System

We explained the attribute selection detailed in previous section. With the information we obtained, this section lists the procedure of our approach in Figure 3, including data preprocessing, attributes model, parameter selection and SVM algorithm.

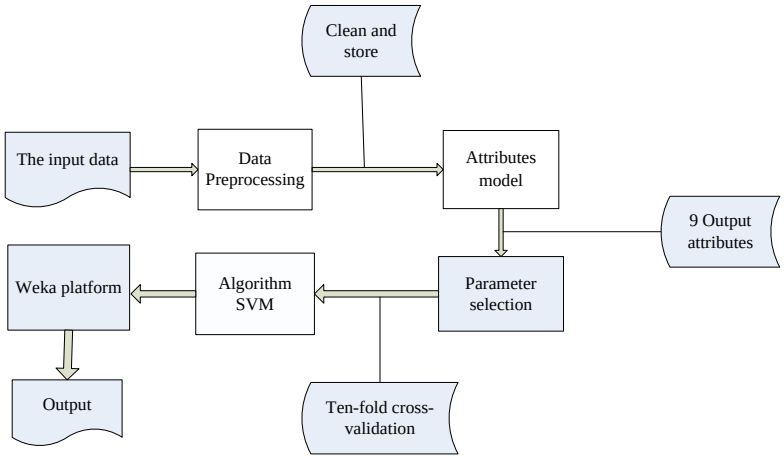

Figure 3. The flow-process diagram of recommender system

Firstly, the raw data provided by Tencent Weibo must be processed to remove illegal data format and useless data. Second, establish a database to facilitate attribute extraction. Then nine attributes are abstracted from the training set and stored in the database. Later, choose the appropriate properties for SVM algorithm. Finally, perform the classification problem using SVM.

3.1. Data preprocessing

Data preprocessing stage is an important part of the total work. The particular issue we need to address in this study is that the samples in the training data are unbalanced. Approximately, 93% of users rejected or ignored, only 7% of users accepted.

We often meet the problem that the distribution of samples is unbalanced. If we don't do anything, it will affect the result model of classification. The general method is to gain an approximate balance of two types of samples by increasing dummy sample to the minority class [Deng and Tian 2004]. With all these methods, reducing negative samples to achieve the approximate value of 1 is the more suitable approach for us.

Besides, we can learn the user's log frequency indirectly through the offered data. They give all of the log information for three months. There are a lot of inactive users in the user dataset. If we take them into consideration in the training set, the accuracy will be greatly reduced. For this reason, we removed the users who use the microblog less than 15 times a month.

4. Case Study

The first experiment focuses on the performance of SVM algorithm on the friend recommendation. the performance of SVM classifier is compared with two classic classification algorithms, Naïve Bayers and Random Forest. The aim of the second experiment is discovering the most influence factor in the user's decision making.

4.1. Friend Recommendation by SVM

For measuring the performance of proposed model in different situations, we construct a series of training data from the original dataset. Every training data has a different size of samples. To guarantee the quality of data, we selected the samples randomly. The twelve sets are consisting of (1 000, 5 000, 10 000, 20 000… 100 000) samples.

There are two reasons for selecting these twelve data sets. The first is preventing errors. The second reason is that choosing different sizes enable to test the performance of SVM classifier in different scale of datasets.

It can be concluded that the SVM algorithm shows a consistent performance for different sample sizes. the F-measure values range from 0.706 to 0.752 and Correctly Classified Instances(CCI) ranges from 0.712 to 0.756. Therefore, the largest variation is 0.046, which is considered a very small variation.

Table 2. Comparison with the SVM

Algorithm	precision	Recall	F-measure	CCI
Naivebayes	↑2.71 %	↑2.8 %	↑2.7 %	↑2.65%
Randomforest	↑7.1 %	↑5.2 %	↑6.0 %	↑5.15%

Table 2 presents the improvement that SVM algorithm has relative to Naïve bayes and Random Forest in Accuracy, Sensitivity, F-measure and CCI.

4.1. The Attribute with Biggest Influence

The importance of this experiment is to identify which attribute actually has the biggest influence. The objective is to select as few attributes as possible to get a good performance.From this experiment, the result discussion is listed as:

a) The training model is particularly sensitive to the acceptance of the followers. In the cases where the attribute "similarity keywords" , "friends in common", "activity" or "acceptance followed" have been removed, there was also an influence on the outcome. Other attributes have a low influence.

b) If the experiment don't use the attributes "acceptance of follower", "key word similarity", "common friends" and "activity", the prediction will become worse. The accuracy may be less than 55%. Therefore, the selected attributes are very reasonable.

5. Future work

The recommender system of this study still leaves room for some improvements: (1) The variation of time for users: This paper did not consider the changes of interest that occur over time; (2) Population Agglomeration: this work did not consider interest among different groups of age or gender. In the future it would be interesting to create another module to consider these characteristics;

References

Swamynathan, G., Wilson, C., Boe, B., Almeroth, K., & Zhao, B. Y. (2008, August). Do social networks improve e-commerce?: a study on social marketplaces. In Proceedings of the first workshop on Online social networks (pp. 1-6). ACM.

Ma, H., Zhou, D., Liu, C., Lyu, M. R., & King, I. (2011, February). Recommender systems with social regularization. In Proceedings of the fourth ACM international conference on Web search and data mining (pp. 287-296). ACM.

Ting, I. H., Chang, P. S., & Wang, S. L. (2012). Understanding Microblog Users for Social Recommendation Based on Social Networks Analysis. J. UCS, 18(4), 554-576.

Tian, D., & Zheng, W. M. (2012, December). Chinese Microblogger character analysis using SVM. In Wavelet Active Media Technology and Information Processing (ICWAMTIP), 2012 International Conference on (pp. 385-389). IEEE.

Sandes, E., Weigang, L and Melo, A. (2012) "Logical model of relationship for online social networks and performance optimizing of queries". In Proceedings of Web Information Systems Engineering - WISE 2012, X.S. Wang, I. Cruz, A. Delis, et al. Springer Berlin Heidelberg, LNCS: 726–736.

Li, X., Guo, L., & Zhao, Y. E. (2008, April). Tag-based social interest discovery. In Proceedings of the 17th international conference on World Wide Web (pp. 675-684). ACM.

Shechtman, O. (2013). The Coefficient of Variation as an Index of Measurement Reliability. In Methods of Clinical Epidemiology (pp. 39-49). Springer Berlin Heidelberg.

Deng, N. Y., & Tian, Y. J. (2004). A new method of data mining: support vector machines. Science Publication, Beijing City.