CSCanada
International Business and Management
Vol. 7, No. 1, 2013, pp. 23-28
DOI:10.3968/j.ibm.1923842820130701.1100

ISSN 1923-841X [Print]
ISSN 1923-8428 [Online]
www.cscanada.net
www.cscanada.org

User Follow Prediction of Microblog Based on the Activeness and Interest Similarity

CAO Yunzhong[a],*; SHAO Peiji[b]; LI Liangqiang[c]; ZHOU Kairui[d]

[a] School of Management and Economics, University of Electronic Science and Technology of China, 610054, Chengdu, China; Business School, Sichuan Agricultural University, 611830, Chengdu, China.
[b] School of Management and Economics, University of Electronic Science and Technology of China, 610054, Chengdu, China.
[c] School of Management and Economics, University of Electronic Science and Technology of China, 610054, Chengdu, China.
[d] College of Economics and Management, Sichuan Agricultural University, 611130, Chengdu, China.
* Corresponding author.

Received 13 May 2013; accepted 16 July 2013.

Abstract

Many microblog fans have formed the basis of microblog information dissemination and diffusion, so accurate prediction and attracting more potential customers to follow microblogger becomes very necessary. By employing interpersonal relationship network of microblog fans, this study integrates microblogger popularity and user activeness into interest similarity in order to explain user follow predictor and propose user follow prediction models. Support vector machine (SVM) is used to train this model. Open data from Tencent microblog in KDD Cup 2012 prove that the proposed prediction model has higher prediction accuracy and stability.

Key words: Microblog; User follow; Microblogger popularity; User activeness; Interest similarity; Prediction model

CAO Yunzhong, SHAO Peiji, LI Liangqiang, ZHOU Kairui (2013). User Follow Prediction of Microblog Based on the Activeness and Interest Similarity. *International Business and Management, 7*(1), 23-28. Available from: http://www.cscanada.net/index.php/ibm/article/view/j.ibm.1923842820130701.1100
DOI: http://dx.doi.org/10.3968/j.ibm.1923842820130701.1100

INTRODUCTION

Microblog, as an important kind of online social networking platform has become an important means of communication, exchange and sharing knowledge in virtual environment (McFedries, 2007; Ebner, 2008). With the rapid development of microblog services, enterprise has begun to develop marketing activities using microblog platform. On the one hand, the enterprise attracts users' follow, comment and retweet, and converts some of its followers to consumers via Microblog post information about products or services. On the other hand, convenient and effective collection of the users' comments help to learn the users' feedback in time and to grasp their interest and potential demand (Hsu, 2010). Information diffusion in microblog platform not only depends on such factors as the microblogger popularity, the quality and timing of released information, but also relies on the participation in comments and retweets. Therefore, by predicting what users will focus on microblogger, attracting the potential microblog fans to follow the microblogger becomes very important. By employing interpersonal relationship network of microblog fans, this study integrates microblogger popularity and user activeness into Interest similarity to propose user follow prediction model. Support vector machine (SVM) is used to train the proposed model.

The paper is structured as follows. The second part of the paper reviews related work. The user follow predictor and prediction model are proposed in the third part. The fourth part testifies the feasibility of the predictor and accuracy of the prediction model by using actual data. Eventually we summarize the conclusion and present the future work.

1. RELATED WORKS

With the widely application of the microblog, more and more researchers have been engaged in microblog related research. There are three aspects related to this research: microblog adoption, continuance use and influence of

microblog. In adoption research, Guardia *et al.* (2011) established the microblog adoption model (μBTAM) by integrating with such factors as subjective norms, image, perceived usefulness, perceived usefulness, and behaviour intention. It confirmed that the TAM model can be used for modelling microblog adoption. Gunther *et al.* (2009) proposed the theoretical models of the enterprise staff adoption of the internal microblog by expanding the UTAUT model via conducting focus group interview and introducing such factors as the information privacy concern, prestige, ratio of signal-to-noise. Schoendienst *et al.* (2011) found that performance expectations and privacy concerns have significant positive effect on behaviour intention by empirical research.

Microblog continuance use model was proposed by employing information systems continuance model and habit in order to explore antecedents of participation in interaction with microbloggers (Barnes & Böhringer, 2011). They found that satisfaction was affected by perceived usefulness and expectation confirmation. Zhao explored the factors affecting users' satisfaction and continuance intention of microblogging. This study proposed that user satisfaction toward microblog service was positively affected by perceived interactivity, ease of use, telepresence, enjoyment and intimacy. It confirmed that microblog continuance use was significantly affected by such factors as perceived usefulness, satisfaction and habits (Zhao & Lu, 2010).

Along with the development of the recommendation system, recommending microbloggers to users and recommending active users to microbloggers is also experiencing good development. On the one hand, in order to expand information dissemination, microbloggers need to attract more users' follow and keep interactions to improve its activity. Such studies focus on recommending users to specific microbloggers. On the other hand, in the vast information, recommending interesting microbloggers and useful information to users is also an important research direction. The above two types of research are of great significance both in theory and application. Calculation of user model and the semantic similarity of Microblog referring users to implement, Mei implemented recommendation by similarity between user model and semantic microblog (Mei *et al.*, 2012). In order to identify microblog users providing useful information, Armentano proposed an algorithm, based on a user PageRank, to implement user prediction (Armentano *et al.*, 2012).

In addition, there are some researchers studied this problem in social networks from the aspects of demographic characteristics, keywords, tags, and the interaction between the users (Zhao, 2012; Chen *et al.*, 2012; Chen & Tang, 2012). In such studies multi-layer factor decomposition model is used for user modeling and forecasting potential fans, but this method used excessive properties to lead to more computational complexity.

Whether users accept and follow specific microbloggers is associated with the popularity and the user's activeness of microbloggers, and also related with interest similarity between them. Based on the above research conclusion, user follow prediction model is proposed to identify potential followers, and then motivate them to become followers by active concerns.

2. RELATED DEFINITION AND USER FOLLOW PREDICTION MODEL

2.1 Definition of Users' Relationship of Microblog

Relationship between the users in microblog network can be defined by Fans Model (Sandes *et al.*, 2012). Microblog users are known as microbloggers or fans. In the community, if the user A follow B, A is B's follower, and B is A's followee. A and B are called R-friends. Microblog networks is described as a directed graph:$G=(V, E)$, Node in the graph represents user, and the connection between nodes E:$V\times V$, represents the relationship between users. $(A, B) \in E$, show user A follows user B, that is to say A is the follower of user B and B is the followee of A.

Users log in microblog platform, establish contact with microbloggers, and form microblog network through querying, browsing, commenting and rewetting, etc. The main problem is to determine a number of predicted characteristics, and to establish the prediction model of user follow. For convenience, this paper defines a given microblogger as Item, and any other user in the network as User.

2.2 Definition of User Interest

Microblog user interest can predict users' behavior to a certain extent. User interest can be obtained from the text of post or comment, and a series of key words and the weight of user interest can be defined as follows:

$$k_u = \{(e_1, w_1), (e_2, w_2), \cdots (e_m, w_m)\} \quad (1)$$

Among them,e_i represents the i[th] keyword,w_i the weight of the i[th] keyword, m the number of keywords. Data of user interest as shown in Table 1.

Table 1
Data of User Interest

User ID	Interest Vector
1001	<101:0.2; 102:0.3; …>
1002	<101:0.4; 103:0.2; …>
…	…

The weight of keywords can be calculated by using the classic TF-IDF formula, as type 2:

$$w_i = TF(e_i) * \log IDF(e_i) \quad (2)$$

2.3 Predict Feature Selection

Determining prediction feature from microblog interact trait is the key of constructing user prediction model. User

modeling can predict characteristics from multiple angles, such as user activeness and demographic attributes, etc. The aspects of more importance are the microblogger's profile, user activeness and interest similarity between users and microbloggers.

2.3.1 Microblogger Popularity and User Activeness

In topic propagation model, blogger popularity and user activeness are the influence factors of propagation process (Zhao *et al.*, 2009). In the microblog network, the main property index to depict the Item include the number of fans, the number of "AT" (@), the total number of retweets and total number of comments, etc. The more numbers or times of the previous factors, the more popular is the microblogger.

At the same time, the number of microblog post, the number of "AT", the total number of retweets and the total number of comments are also the important indicators of user activeness. User think that participation in the microblogging can meet the purpose of information communication and exchange, and the more active of its retweets and comments, the higher is the activeness. Robert pointed out that the interaction is an important factor to keep users to use microblog (Robert & Pongsajapan, 2009). Meanwhile, the number of "AT" reflects relationship between microblog users and describes the interaction between them. Popularity and activeness constitute the necessary premise of the users' following of the Item. Once the Item with high popularity is recommended to high active user, the latter tend to accept and follow the former. Therefore, this paper will take popularity and activeness as a predictor.

2.3.2 Interest Similarity

In the era of micro propagation, quality of microblog content is associated with the user's following. The more humorous and interesting the content is, the more satisfactory and attractive is the content to the users. Therefore, users' following of the item is not only associated with popularity and activeness, but also associated with the similarity of interest.

In recent years, previous research has shown that the rate of microbloggers' being followed and retweeted is based on the microblog content and the basic characteristics of microblog content. From microblog content (posts and comments) the users' interest can be excavated. In the context of this study, interest similarity is the important factor of the microbloggers' being followed. Therefore, interest similarity can be used as prediction index of following.

2.4 User Follow Prediction Model

On the microblog platform, Item and User's releasing of the information, and user's participating in commenting, retweeting will produce some microblog content, and the repeated word can represent their interest. Through natural language processing techniques, keywords can be extracted from the content of microblogs. User interest model is generated from keyword weight by using Vector Space Model (VSM). Keyword vector generation process is shown in Figure 1.

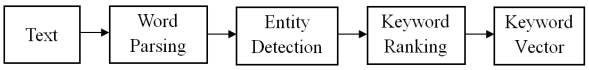

Figure 1
Generation Process of Keyword Vector

After getting interest keyword vector of microblogger u and user v, its interest similarity can be calculated using formula of cosine similarity. As shown in formula 3.

$$sim(u,v) = \frac{k_u \cdot k_v}{\sqrt{\|k_u\|\|k_v\|}} \tag{3}$$

Based on above analysis, the user follow prediction model is proposed. As shown in Figure 2:

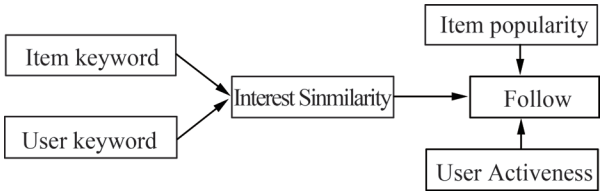

Figure 2
User Follow Prediction Model

3. THE EMPIRICAL RESEARCH AND ANALYSIS

3.1 Data description

To test the feasibility of characteristics of prediction of microblog user follow, this paper uses public datasets from KDD Cup 2012 provided by the tencent Microblog Track 1. The datasets contain seven text file complex number, containing information of items, users and their mutual relationship.In addition, the training set file contains over 70 million records of 31 days. Each record is composed by 'USER', 'ITEM', and 'RESULT'. The meaning of 'USER' and 'ITEM' are stated earlier, the value of 'RESULT' equal 1,means following ,-1, means rejection. Test datasets, containing 19 million records, are used to verify prediction accuracy. Predictors selected in this paper involves five files, the file name and data description are listed in Table 2.

Table 2
Describes Experimental Data

NO.	File Name	Data Description
1	user_profile.txt	Each line includes the following information of a user: year of birth, gender, number of microblog information, the tag IDs.
2	user_action.txt	Each row of data includes the number of "AT" , retweets and comments.
3	user_key_word.txt	Contains each user's keyword and weight value extracted from posts, retweets and comments.
4	user_sns.txt	Contains history of each user's follow information.
5	rec_log_train.txt	Contains three data: user, item, result. The result represents when recommend the item to the user, whether the latter will follow the former.

3.2 Data Preprocess

When choosing training and test data, It has to be guaranteed there is relationship of retweet or comment between item and user, while maintaining their interest similarity is not 0. In this paper, the data preprocessing steps are as follows:

(1) Extract training records from file 5 in Table 2. Sampling is based on the premise that for every pair of user and item, retrieve it if it appeares in same record of the file 2. If any, store the record of the file 5 in the new file. In this way, each record includes interactive relationship between the user and item.

(2) The original training dataset in file 5 has more duplicate records. After above processes, all duplicate records are removed 1.

(3) Taking the user generated in step 2 as indicator, and extracting the number of microblogs in file 1, the total number of 'AT' of the same user, the total number of retweets, and the total number of comments in file 2. All the four attributes describe the user's activeness.

(4) Taking the item generated in step 2 as indicator, extracting number of fans of the item in file 4, the number of 'AT', total number of retweets and total number of comments from other users in file 2, these four attributes describe the popularity of the item.

(5) Taking every pair of user and item generated in step (2) as indicator, extract respectively their interest ID and the corresponding weights in file 3. Interest similarity between each pair of user and item is calculated according to formula 3, eventually, the training dataset is achieved after getting rid of all records with zero score of interest similarity.

After data processing of steps 1-5, 18109 records are produced. The training data are shown in Table 3. In training dataset, 3020 users of all follow the given items.

Table 3
Predictors and its Values

NO.	UserID	ItemID	Follow or Not	Interest Similarity	User Activeness Indicator				Item Popularity Indicator			
					Num. of Posts	Num. of 'AT'	Num. of Retweets	Num. of Comments	Num. of Follower	Num.of 'AT'ed	Num.of Retweeted	Num. of Commented
1	1017156	647356	-1	0.3853	363	0	990	1	296335	17710	135349	27677
2	1017174	2188303	1	0.0353	72	0	8	5	249554	9409	432334	22166
3	1022909	1774797	-1	0.0938	95	0	371	2	89969	2145	291804	11369
4	199979	1606607	1	0.1959	692	10	131	57	5383	1792	19414	2447
5	403208	1791518	1	0.2408	505	7	66	2	27685	863	615294	7891

3.3 The Empirical Results and Analysis

3.3.1 Support Vector Machine (SVM)

Standard support vector machine (SVM), based on structural risk minimization principle ,which is put forward by Vapnik in the 1990s (Cortes & Vapnik, 1995), is a statistical learning method. SVM not only considers the training sample of empirical risk, but also takes into account the generalization ability of the algorithm, so as to effectively improve the accuracy of the algorithm with other machine learning method based on empirical risk minimization principle. SVM, by maximizing the interval of two support planes, realizes the structural risk minimization principle. It transforms a classification or regression problem into a quadratic convex programming problem, so as to make the classifier obtain the global optimal. In the case of nonlinear separation, the nonlinear mapping can be used to map samples of low-dimensional input space to high dimension linear feature space. Therefore, the linear algorithm can be employed to analyse the nonlinear characteristics of the sample in high dimensional feature space. In microblog network, items are recommended to users, these users can choose whether to follow the items or not. Therefore, user follow predict can be transformed into a classification issue, and because the SVM has good classification ability, the SVM is used to complete the prediction of user follow.

3.3.2 Prediction Results and Analysis

In order to test validity of the prediction feature of user follow, this study need to constitute test dataset, and take the way of sampling produce training and test dataset. To avoid different value among attributes influence the prediction results, it needs to normalize such properties as interest similarity, popularity and activeness

In order to avoid error caused by nonrandom sampling, 10 times of experiments are carried out using LibSVM software package, Gauss kernel function, which is used

successfully in many application areas, is selected as the kernel function:

$$K(x, y) = \exp(\frac{-\|x - y\|^2}{2\sigma^2}) \quad (4)$$

Using the grid.py program provided by LibSVM software package to find the optimal parameters of each group of training se γ and C . Use three kinds of prediction precision value P_1, $P_2$and P_{ave} as evaluation indicator. Difinition is as follows $P_i = \frac{|S_i| - |T_{ij}|}{|S_i|}$

$$P_{ave} = \frac{|\bigcup S_i| - \sum_{i,j} |T_{ij}|}{|\bigcup S_i|} \quad i, j = 1, 2, i \neq j \quad (5)$$

Among them, P_1 represents rate by SVM classfying correctly in test dataset composed of records of following the recommended item (that is, the category labeled "1"). P_2 represents rate by SVM classfying correctly in the test dataset composed of records of refusing following the recommended item (that is, SVM also considers the use refusing the recommendation). P_{ave} represents classification accuracy of two kinds of samples. S_i is the i th category test sample collection, T_{ij} is the collection in which i^{th} category samples were divided into j^{th} category by mistake. The experimental results are shown in Table 4.

Table 4
Experimental Results

No.	Penalty Term C	Gamma	Category 1 Accuracy	Category 2 Accuracy	Overall Accuracy
1	32768	0.0078	0.7804	0.3863	0.5833
2	8192	0.125	0.7417	0.3771	0.5594
3	8192	0.0313	0.7031	0.3966	0.5499
4	32768	0.0078	0.7049	0.4043	0.5546
5	512	0.0313	0.6523	0.4614	0.5568
6	8192	0.0313	0.5497	0.5817	0.5657
7	32768	0.002	0.6122	0.4805	0.5464
8	0.03125	0.125	0.1089	0.9194	0.5142
9	8192	0.125	0.6479	0.4798	0.5638
10	8192	2.0	0.5266	0.5681	0.5473
Average Accuracy			0.6028	0.5055	0.5541
Standard Deviation of Accuracy			0.1910	0.1623	0.0177

The findings shown from experimental results in Table 4:

(1) Except experiment 8, the experiment accuracy of the two types of sample is similar. Every experiment achieves the approximate prediction accuracy. However, experiment 8 gets a low prediction precision in the first sample, while at the same time, the prediction is high for the second sample. But the total accuracy of other experiments are still very close. Through the analysis, the causes of this deviation may be that the sample set contains more isolated points. It can be improved by using fuzzy SVM in the future.

(2) All kinds of precision standard deviation of the sample and the total sample pay little, showing that the multiple sampling has little affect to prediction accuracy, and the stability of this prediction method is verified.

Using this method to establish follow predictor, the prediction precision of the 0.5541 is achieved by using support vector machine (SVM). The prediction accuracy obtained by proposed model is slightly better than the previous literatures (Zhao, 2012; Chen *et al.*, 2012; Chen & Tang, 2012), and the selection of indicators and the prediction method is relatively simple and easy to implement.

CONCLUSION AND FUTURE WORK

The increase of active followers to microbloggers help the dissemination and diffusion of microblog information. For microbloggers, by active communication with active users, converting potential active users to their followers to increase popularity and influence is very impotant. User follow predictor and prediction model are proposed in this paper on the basis of research of information dissemination. By using data from KDD Cup 2012 Track 1, and the support vector machine (SVM) for empirical research, This paper achieves a better prediction precision, and comfirms that the proposed indicators such as popularity, activeness and interest similarity are appropriate for predicting user follow.

There are still some shortages in this study. for example, other features such as demographic characteristics, category and tag attributes, etc., may be used for user modeling. In addition, some isolated data points are found in experimental samples, and these isolated points will impact prediction precision of SVM. Therefore, the choice of predictor and classification method still needs to be studied in further research.

REFERENCES

Armentano, M. G., Godoy, D., & Amandi, A. (2012). Topology-Based Recommendation of Users in Micro-Blogging Communities. *Journal of Computer Science and Technology*, *27*(3), 624-634.

Barnes, S. J., Böhringer, M. (2011). Modeling Use Continuance Behavior in Microblogging Services: The Case of Twitter. *Journal of Computer Information Systems*, *51*(4), 1-10.

Chen, T. Q., & Tang, L. P. (2012). Combining Factorization Model and Additive Forest for Collaborative Followee Recommendation. http://www.kddcup2012.org/workshop.

Chen, Y. W., & Liu, Z. T., & Ji, D. Q.(2012). Context-aware ensemble of multifaceted factorization models for recommendation Prediction in Social Networks. http://www.kddcup2012.org/workshop.

Cortes, C., & Vapnik, V. N. (1995). Support-Vector Networks. *Machine Learning*, *20*(3), 273-297.

Ebner, M., & Schiefner, M. (2008). *Microblogging - more than Fun.* In Proceeding of the IADIS Mobile Learning Conference. Algarve, Portugal, 155-159.

Guardia, F. R., Liebana, F. J., & Fiestas, M. M. (2011). *Motivational Factors that Influence the Acceptance of Microblogging Social Networks: The μBTAM Model*. Proc 3rd Int Conf on Education and New Learning Technologies, Barcelona, Spain: IATED, 3196-3206.

Günther, O., Krasnova, H., Riehle, D., & Schöndienst, V. (2009). *Modeling Microblogging Adoption in the Enterprise* (p.544). Proc 15th Americas Conf on Information Systems, San Francisco, California.

Hsu, C. L., Liu, C. C., & Lee, Y. D. (2010). Effect of Commitment and Trust Towards Micro-Blogs on Consumer Behavioral Intention: A Relationship Marketing Perspective. *International Journal of Electronic Business Management*, *8*(4), 292-303.

McFedries, P. (2007). Technically Speaking: All a-Twitter. *IEEE Spectrum*, *44*(10), 84-84.

Mei, K. B., Zhang, B. F., Zheng, J. X., Zhang, L. X., & Wang, M. (2012). *Method of Recommend Microblogging Based on User Model* (pp.1056-1060). IEEE 12th International Conference on Computer and Information Technology.

Robert, A., & Pongsajapan, B. A. (2009). *Liminal Entities: Identity Governance and Organizations on Twitter.* Washington: Georgetown University.

Sandes, E., Li, W. G., & Mello, A. (2012). *Logical Model of Relationship for Online Social Networks and Performance Optimizing of Queries* (726-736). Proceedings of the 13th International Conference on Web Information System Engineering Championship on T1: Scalability.

Schoendienst, V., Krasnova, H., & Günther, O. (2011). *Micro-Blogging Adopting in the Enterprise: An Empirical Analysis* (pp.931-940). Proc 10th Int Conf on Wirtschaftsinformatik, Zurich, witzerland.

Zhao, L., & Lu, Y. (2010). *Perceived Interactivity: Exploring Factors Affecting Micro-Blogging Service Satisfaction and Continuance Intention.* Proc PACIS 2010 on Web 2.0, Social Computing and Data Mining, 940-949.

Zhao, L., Yuan R. X., Guan, X. H., & Jia, Q. S. (2009). Bursty Propagation Model for Incidental Events in Blog Networks. *Journal of Software*, *20*(5), 1384-1392.

Zhao, X. (2012). Scorecard with latent Factor Models for User Follow Prediction Problem. http://www.kddcup2012. org /workshop.

