A fast prediction algorithm for sina weibo users with time correlation cognition

Xi-Xu He, Lei-Ting Chen and Min Zhang
School of Computer Science and Engineering, University of Electronic Science and Technology of China
Provincial Key Laboratory of Digital Media, Chengdu, 610054, China
E-mail: HL@uestc.edu.cn

Social network is the most important way of information exchange and communication at present, which is a key step to study users' behavior in social network. Current methods for users 'behavior classification are more diverse, but it is difficult to assess the impact of specific events in different time. In this paper, the rapid predicted classification algorithm based on correlation time cognition for weibo users in specific events is proposed, which can accomplish weibo user classification according to their behavior in the short time window, so as to establish a more accurate users behavior relation network. Finally, through the experiments, it is shown that the proposed algorithm offers more powerful and robust performance than competing algorithms.

Keywords: Time Related Cognition; Classification Prediction; Granular Computing.

1. Introduction

Many Web social networks such as weibo has promoted some events of network. Many individuals and organizations are using weibo which not only gradually become an important channel for many users to understand the dynamic news, but also become the main source of information for the correct understanding of the user's participation, role and attitude of events of the network.

Weibo users' behavior is defined by web social network, and users publish their speech in accordance with the rules. For example, Sina weibo users just search, comment, transmit and reply information. Users can update their personal information through additions and modifications. System can classify users into ordinary users, the official weibo, membership and the real name authentication users according to their different information. In addition, users can establish a relationship with someone through the "attention" and limit some social functions by setting the privacy.

The way of division above is not suitable for evaluation of users' behavior and construction of users' behavior network. This paper focuses on how to

exploit the users 'behavior to predict its effect in the specific events, so as to realize the accurate construction of users' behavior relationship network.

2. The Analysis of Users Influence and Users Classification

There are many factors related to user influence and many various reflection ways. Due to the impact of the analysis of user influence and the importance of online social networks, many research results, involving the establishment of model and analysis of indirect or external influence and the understanding of relationship between information diffusion and the impact. Tang et.al proposed Topical Affinity Propagation (TAP) method to model social influence based on topic hierarchy in large-scale social network and the experiments on three different data sets show that their method can effectively detect social influence based on the theme [1]. Cui et.al proposed a hybrid factor Non-Negative matrix fact HF-NMF to model social influence of specific matters levels, trying to answer released what content to achieve higher influence through this model [2]. Myers and Shuai et.al analyzed external and indirect influence in Twitter [3][4].

Other scholars also have done many researches on ranking or quantitative method for user influence in Web social network. Kwake et.al ranked the Twitter user influence based on three standards including the number of fans, PageRank and the number of reproduction and they found that ranking results gotten by the number of reproduction is different from ranking results gotten by other two standards [5]. Cha et.al adopted the number of fans, the number of reproduction and the number of user mention, and found that users with more number of fans do not have the more number of reproduction and mention [6].

According to the character and influence of users in the weibo's event, they can be divided into ordinary users (normal) and continuous promoters (special).

Ordinary users have little effect on the scale and development of event, so their comments are rarely transmitted, commented or replied. The continuous promoters who are shown in the development process of event play the role of opinion leader or communication center node in the process of event evolution are more concerned. Because weibo users will not concern about every event in the network, usually just participate in their interested topics.

In some applications, it is very important to quickly identify continuous promoters for the construction of the user relationship network and the mastery of the situation. But the users in weibo event with great random discrete characteristics in time, so according to the temporal characteristics of users 'behavior, the cognize behavior is cognized and the precise division is completed.

3. Data Acquisition Scheme

The public data available on Sina Weibo is seen as the research object, which is captured through the improved web crawler technology and API interface, as shown in Figure 1, the captured data includes weibo event and user information.

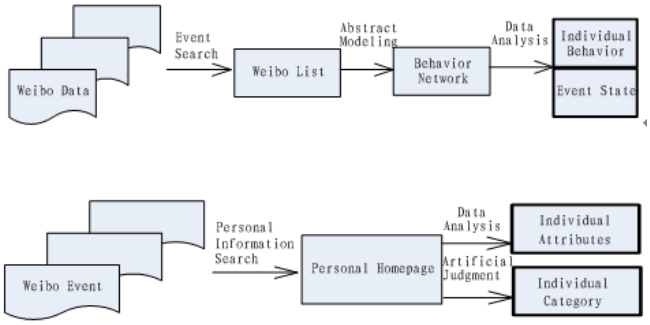

Fig. 1 Data collection structure and Crawl program

Weibo event capture process: firstly, a list of hot events within a month are caught by Sina weibo search function and part of the weibo and relevant data are chosen by artificial selection; secondly, weibo contents are analyzed by the network behavior analysis method to obtain individual behavior of users in the weibo event in and some weibo state parameters, which can be used as part parameters of users behavior features by appropriate treatment; finally the users, whose comment are transmitted by more than fifth other users, can be found out in the weibo event.

Personal attribute data capture process: The information of user is selected by weibo event capture process in their homepage by searching. All personal information and latest weibo (less than 600) are accessed by web crawling technology. And corresponding personal attribute information is obtained by analysis as part parameters of personal behavior characteristic.

4. Classification Algorithm for User's Cognitive with Time

Weibo users are divided into continuous promoters and ordinary users in specific weibo event. Due to time randomness of user participation, this paper proposes cognitive user analysis based on computational analysis algorithm.

Granular computing has marked the emergence of an applied research area involving multi-subjects. The description of granular computing is derived by American mathematician Zadeh, but it is difficult to give a precise and widely accepted definition of granular computing. Zadeh's greatest achievement in granular computing is that he proposed three basic concepts: particle, organization, and causal relationships [7]. Granulation includes the

decomposition from the whole to the part, and the organization includes the integration from part to the whole and the causal relationships include the relationship between the cause and the result. Granular computing does not focus on how to calculate granules and how to apply it directly in specific issues, but rather to the size of the problem.

The definition of granules: set a relationship between domains $R:U \to P(U), \Rightarrow U = \bigcup_{i \in \tau} G_i$ where G_i is called as information particles, and $\{G_i\}_{i \in \tau}$ is granule on domain. The definition above forms the basic particle in granular computing which is the most basic element of granular computing model, and is used to reflect the degree of granulation by granules [8]. The particle is formed by the individual set described by inner attributes and the external attributes. Particle layer is defined by the abstract description for problem space and computing samples. The particle layer of inner particle with same (or similar) granulation or property is constructed by all particles caught by granulation criterion based on some requirement [9]. Granulation is originated from people's perception of the world. Both the measurement of different problems from different angles and the observation, definition and transformation of actual problems belongs to granular computing. The granulation can explain to the size abstraction and complexity in different applications. The relational structure constructed by the relationship among all particle layers is called as Granular structure [10].

The weibo user classification and prediction is actually a process of data mining. Granular computing has been widely used in data mining. The research of literature [11] can be seen that the cluster is carried out under the uniform granulation, and the classification is carried out under different granulation. In this paper, the particle space of the users' behavior characteristics based on the time window is constructed, and the user's network behavior is described as far as the time is concerned.

In accordance with the time window t_i, the attributes of users are classified to form the set $K_i = (U_i, R_i)$ on the time window t_i where any time window attribute subset $X_i \subseteq U_i$ and equivalence relations of classification $R_i \in ind(K_i)$ can obtain cognitive granulation for users' attributes in two time windows:

$$\underline{R}X_i = \bigcup \{x_i \in U_i \mid [x_i]_{R_i} \subseteq X_i\} \tag{1}$$

$$\overline{R}X_i = \{x_i \in U_i \mid [x_i]_{R_i} \cap X_i \neq \varnothing\} \tag{2}$$

According to the formula above, the attributes of different users are formed in accordance with the time window, and then the similarity among user attributes in the event of weibo can be evaluated by computing formula with granular computing, and the user classification can be gotten.

The fifty-eight weibo theme are selected from Sina weibo with social, entertainment, news, finance and other themes and 1,377,734 weibo and 368 users are selected to be analyzed and researched. The three features including "most forms of participation", "most in influence degree" and " the relationship between publisher" selected from the behavior features defined in Table 1 are used to calculate user characteristics under different time window.

4.1. *Most forms of participation*

Users may be involved in weibo many times, so the number of users' participation is one of the criteria for users' classification, where there is a close correlation with the behavior of the users in a time window. The information can be used as an important indicator of user classification.

In Fig. 2, the maximum operation of the user participation mainly includes publication, forwarded and comments, in which the source of weibo comment numbered one and derivative weibo comments numbered four are appeared most. Driven force users want to express their views through participation by weibo comments.

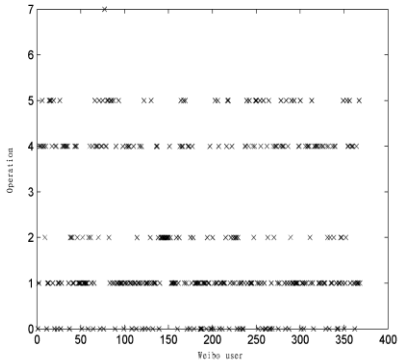

Fig. 2 Most users involved in Weibo operation

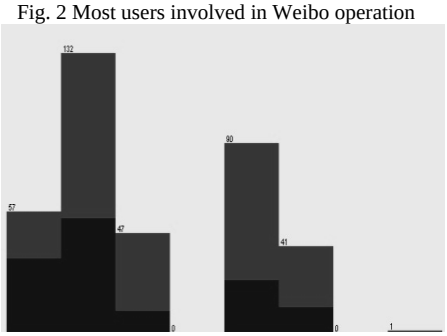

Fig. 3 The relationship between user participation mode and Classification

In Fig. 3, that indicates most publishers will continue to concern on and participate in the weibo in a period of time after the release of weibo. There are least continuous promoters in third operation, indicating that users simply forward to weibo, with lack of sustained participation.

4.2. *Most in influence degree*

Because weibo is an interactive platform between publishers and other followers, other users' participation will often trigger the next action of weibo publisher. Figure 4 shows that the maximum influence concentrates from 100 to 1000, and users with larger influence (more than 10000) are basically the source of weibo's release.

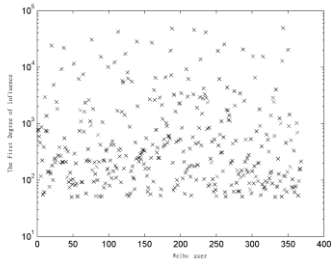

Fig. 4 User participation in the first degree of influence

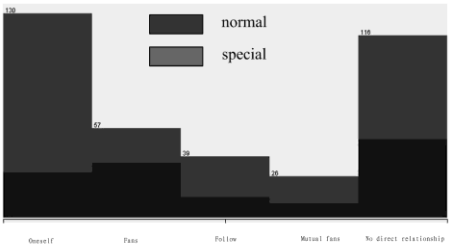

Fig. 5 The relationship between the user and the publisher

4.3. *The relationship between publisher*

In Fig. 2, the highest proportion of continuous promoter (special) is publishers' fans, and the proportion of publishers as continuous promoters is smaller. Compared with complexity of the publisher identity, fan identity is even more single. Different publishers with different identity have different purpose to release weibo. Not all publishers have intended to continue to promote weibo. However, fans concerned about selective news have high subjective will to participate in this weibo. In Figure 5, there is a phenomenon that the portion of users indirect to publisher reaches to 31.5%. This shows that indirect communication is an important way of communication in information dissemination.

5. Experiments

To verify the validity of the proposed algorithm, real data is used to evaluate the performance of the classification. There are many methods to evaluate algorithm with different index. Cross validation is used to evaluate our algorithm in this paper, and the accuracy, recall and AUC (area under ROC curve) and other indicators are used to evaluate the performance of a classifier.

5.1. *Evaluation method*

Holdout collection: data sets D will be divided into two disjoint subsets of the training set D_{train} and test sets D_{test} , where $D = D_{train} \cup D_{test}$ and $D_{train} \cup D_{test} = \varnothing$. The test set is also called as holdout set, that the data is partitioned to D_{train} and D_{train} based on specified proportion.

Cross validation: in k-fold cross-validation, the data set D is randomly divided into disjoint subsets of k , $D_1, D_2, \cdots, D_k$. The amount of data in each subset is roughly equal. Training and testing were carried out k times. In i the iteration, the subset D_i is treated as a test set, and the other subsets are treated as training set. This operation above can ensure that all data are tested.

5.2. *Evaluation index*

The results for each classifier can be established as the matrix with the four indicators defined below (Table 1):

Positive True (TP): the ability to correctly classify the number of ordinary users' characteristics.

Positive False (FP): the number of general user misclassified to the continuous promoter.

Negative True (TN): the number of correct identification of continuous promoter.

Negative False (FN): the number of the continuous promoter misclassified to the general (regular) users

Tab. 1 Mixed matrix of classifier

	Classified as positive	Classified as negative
The actual positive cases	TP	FN
The actual negative cases	FP	TN

The recall, false positive rate and the overall accuracy rate are used to measure classification efficiency of our algorithm, Bagging and random forest, the definition of these three indicators are as follows:

Recall: the proportion of continuous promoters with correct identification in all continuous promoters, as the following formula.

$$DetectionRate = \frac{TP}{TP + FN} \tag{3}$$

False positive rate: the proportion of continuous promoters misclassified to the ordinary users, as the following formula

$$FalseAlarmRate = \frac{FP}{TN + FP} \tag{4}$$

Accuracy: the proportion of ordinary users with correct identification in all users as the following formula.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

In addition, use AUC (Area Under ROC Curve). The ROC curve is widely used in machine learning and data mining to describe the trade-off between recall rate and ales positive rate in one classifier. AUC ($0 \le AUC \le 1$) computes the recall rate through ROC curve calculation. The value of AUC is larger, the TP is higher and the FP is smaller. When $AUC = 1$, the $TP_{rate} = 1$ and $FP_{rate} = 0$.

Tab. 2 shows that the results of our algorithms based on time window user's cognitive classification, bagging algorithm and Forest Random algorithm.

Tab. 2 Using the classification results of the three characteristics

Method	Accuracy(%)	AUC
Classification algorithm for user's cognitive prediction with time	73.9	0.678
Random Forest	73.4	0.797
Bagging	73.9	0.8

6. Conclusion

Information dissemination is resulted by multiple factors, in which different types of users play different roles. In this paper, the characteristics of the Web social network is analyzed based on the method of time window recognition, combined with the formation of granular computing to the user classification prediction, so as to achieve the accurate modeling of user behavior relationship network.

References

1. Jie Tang, Sen Wu,Bo Gao and Yang Wan. Topic-level Social Network Search[C]. Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, p769-772, 2011.
2. Peng Cui,Fei Wang,Shaowei Liu,Mingdong Ou,Shiqiang Yang and Lifeng Sun. Who Should Share What? Item-level Social Influence Prediction for

Users and Posts Ranking[C]. International Acm Sigir Conference on Research and Development in Information Retrieval, p185-194, 2011.

3. SA Myers,C Zhu and J Leskovec. Information Diffusion and External Influence in Networks[C]. Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, p33-41, 2012.
4. X Shuai, Y Ding, J Busemeyer, S Chen, Y Sun and J Tang. Modeling Indirect Influence on Twitter [J]. International Journal on Semantic Web and Information Systems, vol8 (4), p20-36, 2012.
5. H Kwak, C Lee, H Park and S Moon. What is Twitter, a Social Network or a News Media[C]. International Conference on World Wide Web, p591-600, 2010.
6. M Cha, H Haddadi, F Benevenuto and PK Gummadi. Measuring User Influence in Twitter: The Million Follower Fallacy[C]. In Icwsm 10: International Aaai Conference on Weblogs and Social Media, p10-17, 2010.
7. ZADEH L A.Towards a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic [J] .Fuzzy Sets and Systems, vol19, p111-127, 1997.
8. Z Zheng.Tolerance granular space and its applications [D]. Gradua te Schoo l o f the Chinese Acdemy of Sciences, 2005.
9. Wang Guo-Yin, Zhang Qing-Hua , Hu Jun.An overview of granular computing[J].CAAI Transactions on Intelligent Systems, vol2(6), p8-26, 2007.
10. YAO Y Y.Granular computing for data mining [A]. Proceedings of SPIE Conference on Data Mining, 2006.
11. BU Dongbo, BAI Shuo, LI Guojie. Principle of granularityin clustering and classification[J].Chinese Journalof Computers , vol25(8), p810 -816, 2002
12. SHEN Yalan.Research on methods of data mining based on granular computing [D]. Shenyang University of Technology, 2006.