2012 Fourth International Conference on Multimedia Information Networking and Security

Identifying Important Users in Sina Microblog

Jiaqi Liu[1], Zhidong Cao[1], Kainan Cui[1]
[1]State Key Laboratory of Management and Control for Complex Systems
Institute of Automation, Chinese Academy of Sciences
Beijing, China
jiaqiliu.ia@gmail.com, zhidong.cao@ia.ac.cn, kainan.cui@live.cn

Feng Xie[2]
[2]Department of Automation, Tsinghua University
Tsinghua National Lab for Information Science and Technology, Tsinghua University
Beijing, China
xief10@mails.tsinghua.edu.cn

***Abstract*—Important users are high-status vertices in social networks. They are everywhere in most fields of society and have big impact on those around them. Although a lot of effort has been made on identifying important users, the efficient methods still need to be developed, especially for the web users from Sina microblog, which is the most popular social networking sites in China and has unique characteristics. In this paper, a machine learning-based method which only uses several attributes on Naive Bayes Classifiers (NBC) and Back Propagation Neural Network (BPNN) was proposed to identify important users. Initial experiments indicate that our method is effective. The result of "high" category has more than 55% accuracy rate. We find the NBC can identify more important users while BPNN has higher accuracy rate. What's more, the numbers of follower and followings in Sina microblog is independent.**

Keywords-important users; social network; Sina microblog; Naive Bayes Classifiers; Back Propagation Neural Network

I. INTRODUCTION

With the developing of online social network, people surf the Internet looking not only for information, but also for friends. An increasing number of people are using the Web to share their opinions about a wide range of topics ranging from personal relationships, products, services, to political views [1-3]. It is very common for people to read opinions of others and share their views. As time goes on, some nodes of social network came to a high status. They are the opinion learners. Important users are "those individuals to whom others turn for advice and information", they play an important role in farming and reflecting the opinions of the masses [3-10]. Hence, identifying them in social network and monitoring their opinions can be helpful for forecasting the previous hot topic and take relative measures in advance.

Some literatures about identifying important users in online social network have been published, but few referred to Sina microblog, which is the most popular social networking sites in China with unique characteristics. Reference [11] measured and analyzed the structural properties of Orkut, Youtube, Flickr and LiveJournal. They observed that the indegree of user nodes tends to match the outdegree. Reference [12] analyzed the characterization of Twitter using the number of users' followers and that of followings. Reference [13] characterized the structural properties of microblog, such as degree distribution, radius, and reciprocal rate. They proved microblog is different from human social network and other online social network.

In this paper, we proposed a machine learning-based method which only needs several attributes to identify important users in microblog. The question we attempt to answer is how to identify the important users in microblog with part of some basic profiles of users.

The rest of the paper is organized as follows: Theoretical background and experiment design introduced in the next section. Section 3 describes our dataset and preprocessing in detail. The experiment results and corresponding analysis are presented in Section 4. Finally, we draw our conclusions and outline our future work in Section 5.

II. THEORETICAL BACKGROUND AND EXPERIMENT DESIGN

In this section, we will provide theoretical background in advance and then describe our experiment design.

A. Theoretical background

The Naive Bayes Classifier (NBC) and the Back Propagation Neural Network (BPNN) are widely used in pattern classification. Both of them are teacher training network with definite structure which means we do not need to learn the model, but only need to estimate parameter according to the training data. What's more, betweenness centrality is one of measures that widely used in graph theory. I will give a brief introduction of these theories as follow.

1) Naive Bayes Classifier

A Naive Bayes Classifier (NBC) is a simple probabilistic classifier based on applying Bayes' theorem with strong independence assumptions. It has a definite structure which means we don't need to learn model and only need to estimated parameter according to the training data [14].

The Naive Bayes assumption is that within each class, the values of the attributes of examples are independent, as in (1).

$$p(x_1,\cdots,x_n|c)=\prod_{i=1}^{n}p(x_i|c) \quad (1)$$

978-0-7695-4852-4/12 $26.00 © 2012 IEEE
DOI 10.1109/MINES.2012.122

Then, we can apply Bayes rule to compute the posterior

$$p(c|x_1,\cdots,x_n)=\frac{p(c,x_1,\cdots,x_n)}{p(x_1,\cdots,x_n)}=\frac{p(c)p(x_1,\cdots,x_n|c)}{p(x_1,\cdots,x_n)}$$

$$=\frac{p(c)\prod_{i=1}^{n}p(x_i|c)}{p(x_1,\cdots,x_n)}=\alpha p(c)\prod_{i=1}^{n}p(x_i|c) \qquad (2)$$

where α is just standardization constant and do not have relationship with c.

Finally, we use the following function to predict the class of examples.

$$\arg\max_{c(x_1,\cdots,x_n)}\left\{p(c)\prod_{i=1}^{n}p(x_i|c)\right\} \qquad (3)$$

Since using $p(c)\prod_{i=1}^{n}p(x_i|c)$ to approximate joint probability $p(c,x_1,\cdots,x_n)$, it probability make some error.

The advantage of the NBC is that it only requires a small amount of training data to estimate the parameters necessary for classification. Because independent variables are assumed, only the variances of the variables for each class need to be determined and not the entire covariance matrix. Hence, the NBC perhaps the most popular approach to classification.

2) Back Propagation Neural Network

The Back Propagation Neural Network (BPNN) is widely used in pattern classification. In theory, a BP neural network with three layers can solve arbitrary classification. The structure is shown in Fig. 1 [15-17].

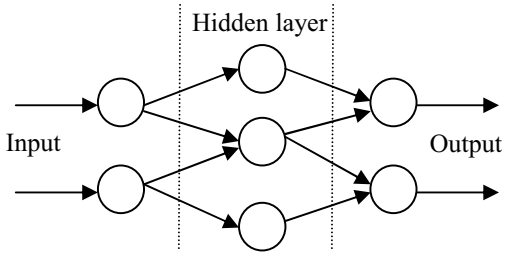

Figure 1 The structure of BPNN

BPNN is a teacher training network, and we must provide a learning set that consists of some input samples and the known-correct output for each case. Its learning process works in small iterative steps: one of the example cases is applied to the network, and the network produces some output based on the current state of its synaptic weights. This output is compared to the known-good output, and a mean-squared error signal is calculated. The error value is then propagated backwards through the network, and small changes are made to the weights in each layer. The weight changes are calculated to reduce the error signal for the case in question. The whole process is repeated for each of the example cases, then back to the first case again, and so on. The cycle is repeated until the overall error value drops below some pre-determined threshold. At this point we say that the network has learned the problem "well enough" - the network will never exactly learn the ideal function, but rather it will asymptotically approach the ideal function [18].

3) Betweenness Centrality

Betweenness centrality was introduced as a measure for quantifying the control of a human on the communications between other humans in a social network by Linton Freeman [19, 20]. In his conception, vertices that have a high probability to occur on a randomly chosen shortest path between two randomly chosen nodes have a high betweenness, which also means that high-betweenness person has strong connection to other people. Hence, we use betweenness centrality to measure the user's importance in our experiment.

B. Experiment design

In order to identify the important users in microblog with part of some basic profiles of users, we collect a dataset and using part of them for training the NBC and the BPNN. The data set is described in detail in next section. Since the number of followers, followings, tweets and "whether verified" for each user have strong relationship with user's importance, we use these attributes as input for training models. Then, we identify the important users in the rest of dataset with the trained models. In theory, we can identify every user in Sina microblog. But we can't examine our identification result since the betweenness centrality is measured in a group people. Therefore, we use the rest data as test set to verify our result. Finally, we analyze the result and draw our conclusion (Fig.2).

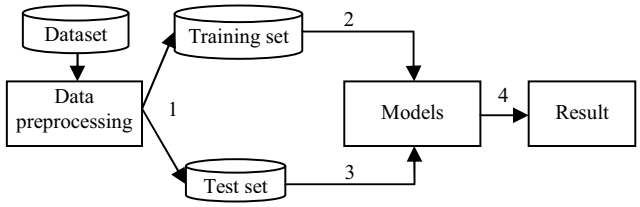

Figure 2 The flow chart of experiment

III. DATASET

A. Dataset

7.23 China Railway High-Speed (CRH) train accident happened at night of July 23, 2011. After the accident, many Sina microblog users commented on the accident and shared their opinions. Voting system of Sina microblog gives us an easy and objective way to identify users' opinions. We can make clear about the users' opinions by just collecting their options. Thus, we choose a vote titled "Will you still support CRH?" The voting started at 15:05, July 26, 2011.

We used the Application Programming Interface (API) provided by Sina microblog to collect data. We collected profiles of users who have voted and each profile includes

user ID, user's full name, location, gender, number of followings, number of followers, number of tweets, relationship of each other, whether verified, and their options as well as the vote order. Some basic statistical information is shown in Fig.3.

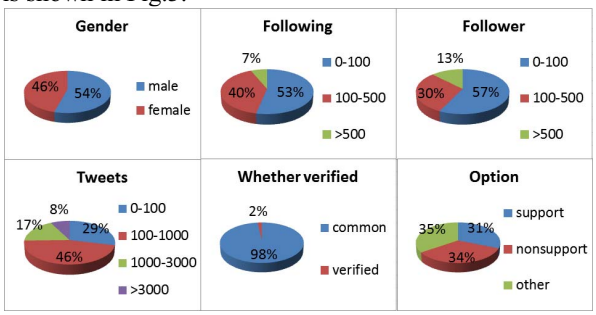

Figure 3 Basic information of dataset

B. *Data preprocessing*

The voting process ran until 15:04, August 2, 2011. Finally, we got 2071 users' profiles and their relationships (3888 edges). We find that 935 users have no access to any other users, which means they are the isolated nodes. The graph of rest users were shown in Fig.4, where maroon dot means nonsupport, green dot means support and black dot means the other. The size of the vertices and the logarithm of vertices' degree are proportional.

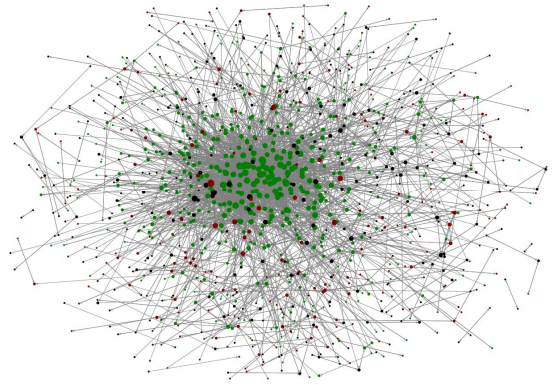

Figure 4 The graph of users

1) Importance classification

We calculated the degree and betweenness centrality of each vertex with NodeXL. Since our dataset was collected based on voting system, we simply assumed that everyone voted will share their opinion via microblog, and if one is the important user, many of his or her followers will follow and vote. According to this assumption, one with more importance has high betweenness centrality.

The Pareto principle, also known as the 80–20 rule, states that, for many events, roughly 80% of the effects come from 20% of the causes. Hence, we classify 20% users with high betweenness centrality in "high" category and the rest 80% in "low" category.

2) Input classification

BPNN is good at classifying data with complex relationships. Therefore, we classify the number of followings and followers separately. On the other hand, NBC has strong independence assumption. The separation is 225 for following as well as 352 for follower. But the number of followings and followers seem to have apparent dependence. In that case, we try to use the "follower/following" as one attribute to eliminate their dependence. The separation is 2.0. Besides, we also use the classification that BPNN used as a comparison.

Reference [13] presents that the number of tweets is not related to the number of user's followers and followings. Thus, we just consider the number of tweets as an independent input, and classify based on Pareto principle. The number of tweets for the top 20% users is more than 1277.

Additionally, user can be classified into two categories: verified user (Vuser) and common user (Cuser).

In summary, we train NBC-1 with data of "follower/following", "tweets", "whether verified" in training set. And we use "followers", "followings", "tweets", "whether verified" for training the NBC-2 and BPNN.

3) Training set and test set

Since NBC and BPNN both need to learn parameter from dataset, we randomly sampled 70% users as training set and the rest as test set. The percentage of two sets' importance classification is almost the same (Table I).

TABLE I. THE OPINION DISTRIBUTION OF EACH CATEGORY IN TRAINING AND TEST SET

Category	Training set		Test set	
	Number	*Percentage (%)*	*Number*	*Percentage (%)*
Total	1450	100	621	100
Low	1170	80.69	486	78.26
High	280	19.31	135	21.74

IV. RESULTS AND DISCUSSIONS

When completing training, we test each method with the data of test set. Table II illustrates the results of number of "high" category.

TABLE II. COMPARISON AMONG RESULTS OF THE THREE METHODS

Method	*Number*	*Correct*	*Accuracy rate (%)*	*Selected rate* [a] *(%)*
NBC-1	65	37	56.92	27.41
NBC-2	116	67	57.76	49.63
BPNN	13	12	92.31	8.89

a. The selected rate means percentage of the correct selection in "high" category.

To our surprise, the accuracy rates of NBC-1 and NBC-2, 56.92% and 57.76% respectively, are very close, and the number of "high" category of NBC-2 is almost double with NBC-1, that means NBC-2 identifies most important users in test set, and it is more effective than other methods. Besides, comparing the different inputs of these two methods, we can say the number of followers and followings are strong independent, which is different from what they appear.

In terms of BPNN, despite only 13 important users were classified, it has very high accuracy. In other words, we can simply use this method to identify some representative important users in a large number of people. Monitoring users of the result list will have big influence on forecasting hot topics.

TABLE III. OPTION DISTRIBUTION IN THREE CLASSIFICATIONS

Method	Low (%)			High (%)		
	Support	*Non support*	*Other*	*Support*	*Non support*	*Other*
Dataset	21.74	39.49	38.77	68.67	13.73	17.59
Test set	24.07	38.48	37.45	72.59	11.85	15.56
NBC -1	32.55	33.45	33.99	52.31	26.15	21.54
NBC-2	31.49	32.67	35.84	48.28	32.76	18.97
BPNN	34.54	32.73	32.73	38.46	30.77	30.77

Moreover, we find in "high" category of test set, a large proportion of users (72.59%) vote "support". But in the "low" category the percentage is different, as shown in Table III. The percentage of "high" category is also higher than others in our NBC based methods. In our classification, those who classified incorrectly in "high" category almost vote "nonsupport", with 75% for NBC-1and 85.71% for NBC-2. It means that users who had a positive mood tended to more important in microblog. We can also find the same phenomenon in Fig.4, where the number of green is far more than red and black in center. In other words, we should pay more attention to whom in positive mood.

There are 13 Vusers in the test set, 12 of them are in the "high" category. NBC-1 made a completely correct classification, where NBC-2 and BPNN only classified one sample in the wrong category. Our results confirmed that "Whether verified" is an important attribute, which is similar to [13]. Besides, some famous people verified in microblog, but don't really important. That's because of the following two reasons. Some Vusers have specific field that they can be important to users. Such as "Shijiazhuang securities exchange" surely has big influence on securities exchanging, but poor in other fields. On the other hand, some Vusers neither share their opinions nor vote. Hence, despite having verified, they have low betweenness centrality.

V. CONCLUSION AND FUTURE WORK

In this paper, we introduce the basic characteristics of microblog and some theoretical background of machine learning, and propose a machine learning-based method which only needs several attributes to identify important users. It's one of the first attempts to identify the important users in Sina microblog with only some basic profiles of users. Our results indicate that identify important users with only little information is possible and the NBC as well as BPNN are good model to be used. We also find the number of followers and followings has strong independence.

In the future, we plan to extract and examine more individual attributes and added in to these models. Such as the change of user's follower, the time of user sharing opinion. We also plan to extend the analysis by using other models of machine learning or statistics.

ACKNOWLEDGMENT

This work was supported in part by the Major-projects of Science and Technology Research (Grant No. 2012ZX10004801) and National Natural Science Foundation of China (Grant No. 90924302, 91024030， 40901219, 71050001).

REFERENCES

[1] S. Bin and C. Kuiyu, "Mining Chinese Reviews" in Proc. of ICDM Workshops, 2006, pp. 585-589.

[2] Cho, K. S., J.-S. Ryu, et al. (2010). "Credibility Evaluation and Results with Leader-Weight in Opinion Mining." 5-8.

[3] Meng, F., J. Wei, et al. (2011). "Study on the Impacts of Opinion Leader in Online Consuming Decision." 140-144.

[4] Zhou, H. M., D. Zeng, et al. (2009). "Finding Leaders from Opinion Networks." Isi: 2009 Ieee International Conference on Intelligence and Security Informatics: 266-268.

[5] Zhongwu Zhai, Hua Xu, Peifa Jia. "Identifying Opinion Leaders in BBS" WI-IAT, 2008, 398-401.

[6] Rogers, E.M., Characteristics of agricultural innovators and other adopter categories. 1961: p. 882.

[7] Rogers, E.M., Diffusion of Innovations. 2003, New York: The Free Press.

[8] X. Song, Y. Chi, K. Hino, and B. Tseng, "Identifying opinion leaders in the blogosphere,"Proceedings of the Conference on Information and Knowledge Management (CIKM), pp. 971–974, 2007.

[9] Nitin Agarwal, Huan Liu, Lei Tang, and Philip S. Yu. Identifying the influential bloggers in a community. In WSDM '08:Proceedings of the international conference on Web search and web data mining, pages 207–218, New York, NY, USA, 2008. ACM.

[10] Yu-tao, M., Shu-qin C.,Rui W. "Study on the method of identifying opinion leaders based on online customer reviews." Management Science and Engineering (ICMSE), 2011 International Conference: 10-17

[11] A Mislove,M Marcon,KP Gummadi,P Druschel,B Bhattacharjee, "Measurement and analysis of online social networks." IMC'07, October 24-26, 2007, San Diego, California, USA.

[12] B. Krishnamurthy, P. Gill, and M. Arlitt. "A few chirps about twitter." In Proc. Of the 1st workshop on Online social networks. ACM, 2008.

[13] Zhengbiao Guo, Zhitang Li, Hao Tu, "Sina Microblog: An Information -driven Online Social Network." CW, 2011,pp,160-167

[14] Kevin P. Murphy, " Naive Bayes classifier", Department of Computer Science, University of British Columbia

[15] Martin T.Hagan, Howard B. Demuth, Mark H.Bcale. Neural Network Design[M]. China Machine Press, Beijing, 2002, pp.227-228.

[16] R.Hecht-Nielsen, "Theory of the Backpropagation Neural Network" in Proc.IEEE-IJCNN89(Washington DC),1989,pp.593-605, vol.1

[17] Jiang, J. F., J. Zhang, et al. (2010). "Application of Back Propagation Neural Network in the Classification of High Resolution

[18] McCollum, P. (1998) An Introduction to Back-Propagation Neural Networks, Encoder. URL: http://www.seattlerobotics.org/encoder/nov98/neural.html

[19] Tore Opsahl, Filip Agneessens, John Skvoretz. "Node centrality in weighted networks: Generalizing degree and shortest paths." Social Networks, 2010, 32 (3): 245-251

[20] Freeman, Linton. "A set of measures of centrality based upon betweenness." Sociometry 1977,40: 35–41.