Predicting User Participation in Social Networking Sites

Qingchao Kong[1] Wenji Mao[1] Daniel Zeng[1,2]
[1]State Key Laboratory of Management and Control for Complex Systems,
Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
[2]Department of Management Information Systems, University of Arizona, Tucson, AZ 85721, USA
{qingchao.kong, wenji.mao, dajun.zeng}@ia.ac.cn

***Abstract*—Social networking sites provide a convenient way for users to participate in discussion groups and communicate with others. While users situate in and enjoy such a social environment, it is important for various security related applications to understand, model and analyze participating users' behavior. In this paper, we make an attempt to model and predict user participation behavior in discussion groups of social networking sites. Our work employs a feature-based approach, which considers four types of features: thread features, content similarity, user behavior and social features. We conduct an empirical study on a popular social networking site in China, Douban.com. The experimental results show the effectiveness of our approach.**

Keywords—user participation; behavior modeling and prediction; social networking sites

I. INTRODUCTION

Social networking sites (SNS), such as Facebook and Twitter, have had enormous impact on people's online activities. People update status or tweet about their life events and feelings, and share information with their friends. Social networking sites have become major platforms for people to express opinions and get involved in discussions. While SNS support various applications, it is important for these applications to understand, model and analyze participating users' behavior. For example, in information retrieval, predicting who will be interested in a thread and participate in the discussions help service providers get more accurate information about the interests and needs of their users. In security-related applications, tracking how discussions evolve over time and predicting who will participate in discussions help decision-makers and security agents better understand and monitor the occurrence of abnormal user activities.

In this paper, we present an approach to predicting user participation behavior in social networking sites. Specifically, we focus on predicting user comment behavior in SNS thread discussions. We propose a feature-based approach which considers several key factors that affect users' comment behavior, such as thread content, user participation history, user interests and relationships among users. We build logistic regression models to acquire probabilities of user participation, which represent how likely each user will comment on a specific thread.

This work has made several contributions. It is among the first attempts to tackle user participation prediction problem in a social network environment. We combine different types of features, such as user activity history and number of common friends, to capture behavioral and social aspects of the prediction problem, and explore the tree structure of user comments.

II. RELATED WORK

User comment behavior is commonly seen in blogs or forums. There has been several research on predicting the volume of comments given a particular blog post. Tsagkias *et al.* [1] examined a set of surface, cumulative, textual, semantic and real-world features to predict whether a blog post will receive any comments and whether the volume of comments is high or low. Yano and Smith [2] used a topic modeling approach to predicting the volume of comments. As predicting the volume of retweets is similar to that of comments: Hong *et. al* [3] proposed a feature-based method based on content of the messages, temporal information, metadata of messages and users, as well as structural properties of the users' social graph.

In addition to predicting the volume of user comments, predicting comment behavior (i.e., predicting who will make comments) is a more challenging task. Yano *et al*. [4] developed a probabilistic model for the generation of blog posts and comments jointly to predict which posts a user will respond to. Tang *et al*. [5] proposed a User Interest and Topic Detection (UTD) model to capture topics and trends and predict who might be interested in the newly published threads, given that some users had already commented on this thread.

These works have addressed a couple of factors that affect user behavior in blogs and forums, especially thread and content information. However, for social networking sites, there are additional social and behavioral factors that can be utilized to predict user behavior. For example, a user may leave a comment just because they are "fans" of the thread publisher.

In this paper, we propose a feature-based method to capture a variety of these key factors. We represent the main features that affect user comment behavior. These features include not only thread and content features, but behavioral and social features as well. We combine multiple features to predict user comments via machine learning methods. The prediction algorithm adopts the widely used logistic regression models to predict the probability of user commenting on a thread, given only the information of thread content and publisher.

III. PROBLEM DEFINITION

We formulate the problem in this section. Let D denote a set of discussion threads and U denote a set of users from a

social networking site. Each thread d ($d \in D$) consists of the title and content, both of which are represented by bag-of-words, and the publisher of this thread u_d ($u_d \in U$). Apart from the content of the thread, we also have the information of social connections, i.e. following information, and activity history of all the users in U.

Assuming for thread d, $d \in D$, the set of users who have truly commented is defined as $S_d = \{u_1, u_2, \cdots, u_n \mid u_i \in U\}$. We also define "target user set" as $TS_d = \{u_1, u_2, \cdots, u_m \mid u_i \in U\}$, where $TS_d \supseteq S_d$. Our problem is defined as: given the target user set TS_d for a thread d, predict the probability of a target user u in TS_d commenting on d.

IV. Proposed Method

In this section, we introduce the features in detail and combine all features altogether to build logistic regression models for prediction.

A. Feature Engineering

There are four types of features in use, namely thread features, content similarity, user behaviors and social features.

1) Thread features: including length of the thread title, length of the thread content, number of images and outlinks in the thread content.

2) Content similarity: including the similarity between the thread content and the target user's interest, and the similarity between the thread publisher's interest and the target user's interest. In order to compute similarities, a LDA [6] model is trained using the whole corpus, which includes all the threads' content and comments. All comment texts and the content of all the threads a user have contributed to, are gathered to form a document, which will then be fed into the LDA model to get the topic distributions as the user's interest. By using the LDA model, all document texts and user interest can be represented as topic distributions, i.e. vectors of specified length. Content similarity is defined as the Euclidean distance between the two topic distributions.

3) User behavior: including the number of threads the thread publisher and the target user have published, the number of comments the thread publisher and the target user have contributed.

4) Social features: including the number of followers the thread publisher and the target user have, the number of followees the thread publisher and the target user have, the number of mutual followees the thread publisher and the target user share, and whether the target user is following the thread publisher.

B. Logistic Regression Model

We use the logistic regression (LR) model combine all the features we have extracted. LR model is a linear classification model and can assign probabilities to each test instance indicating how likely the test instance belongs to each class. For each target user in target user set and each thread, we extract all the features described above to make a feature vector. Specifically, let $\mathbf{x}$ be the feature vector and $\mathbf{w}$ be the weight vector for the corresponding features. Y is a binary random variable which represents the prediction outcomes: the target user will comment ($Y = 1$) and the target user will not comment ($Y = 0$). Let

$$P_{\mathbf{w}}(Y = 1 \mid \mathbf{x}) = \frac{1}{1 + e^{-g(\mathbf{x})}} \tag{1}$$

where $g(\mathbf{x}) = w_1 x_1 + w_2 x_2 + \cdots + w_n x_n$. $P_{\mathbf{w}}(Y = 1 \mid \mathbf{x})$ gives the probability of a target user commenting on the given thread. According to (1), the model predicts $\hat{y}(\mathbf{x}) = 1$ when $g(\mathbf{x}) > 0$. Otherwise, it predicts $\hat{y}(\mathbf{x}) = 0$, when $g(\mathbf{x}) < 0$.

C. Ensemble of LR models

In our problem, the negative instances may be much more than the positive instances, thus causing the class imbalance problem. We define the "class imbalance ratio" as follows:

$$R = \frac{(\#\text{negative} + \#\text{positive}) \text{ instances in training set}}{\#\text{positive instances in training set}} \tag{2}$$

Usually R is chosen to be a positive integer.

In order to solve the class imbalance problem, ensemble methods are used. Specifically, we first equally split the whole negative instance set into $R-1$ sub-negative instance sets, then each sub-negative instance set is combined with the whole positive instance set to make a new training set, so that each new training set is balanced, and we use $R-1$ new training sets to build $R-1$ LR models. Our prediction framework uses a voting mechanism to predict how likely that a user will comment on a certain thread. For one test instance in the test set, all the $R-1$ LR models give their predictions. Assuming that there are n_pos positive votes and n_neg negative votes, the prediction framework computes the probability that a user will comment on a thread as $n\_pos / (n\_pos + n\_neg)$.

V. Experiment

A. Dataset and Preprocessing

We use the Douban Group dataset. Douban.com is a very popular social networking site in China, with about 70 million users. Being part of Douban, Douban Group allows users to form different groups with different interests. Group members can publish a thread and others can comment, recommend the thread to their followers or simply click "Like". We have crawled all the threads and comments in Douban group "ustv".

The training set includes threads and their comments published during the period from August 1st to December 1st, 2012, while the threads in the test set are published during the period from December 1st, 2012 to January 1st, 2013. In the test set, some users and threads are removed to guarantee that every user in the test set has at least 2 comments and every thread in the test set has at least 5 commenting users.

TABLE I shows the statistics of threads, comments and users after preprocessing.

TABLE I STATISTICS OF TRAINING AND TEST SET

	# of threads	# of comments	# of users
Training set	535	20473	3467
Test set	112	4226	708

B. Target User Set

It is natural to treat the set of all users in a group as the "target user set" as defined in Section III, but there are usually several thousand group members in every group (170,149 members in group "ustv" so far), so it is almost impossible to find the users who will comment on a thread among so many users. Thus we use a much smaller target user set.

For each thread d, TS_d is constructed as follows: first add all users in S_d to TS_d so that $TS_d \supseteq S_d$, then randomly select $(R-1)\times|S_d|$ users from other group members who are not in S_d and add them to TS_d, where $|S_d|$ is the number of users in S_d. So for each thread d in the training set, there are $|S_d|$ positive instances and $|TS_d|-|S_d|$ negative instances. The class imbalance problem becomes more difficult as R becomes larger.

C. Evaluation Criterion

Traditional evaluation criteria like precision and recall are not good options to test our method, since the test set may be very unbalanced. User participation prediction can also be viewed as an "information retrieval" problem ("user retrieval" in this case), which is to retrieve users with high probability that they will comment on a certain thread. So we choose Precision@K, which is commonly used in information retrieval. The prediction framework returns the top-K most probable users who will comment on each thread, denoted as the set $topK$, and Precision@K is calculated as:

$$Precision@K_d = \frac{|topK_d \cap S_d|}{K} \quad (3)$$

D. Results

For each R, the experiment is repeated five times and the average Precision@K for each thread is reported. Fig. 1 shows Precision@5 distributions for each thread when R =4, 6, 8 and 10, respectively. Note that to make the figure more readable, all the threads' Precision@5 are sorted in descending order.

As we can see in Fig. 1, for each R, our prediction framework performs well for some threads but very poorly for others(the Precision@5 reaches 0), i.e. the prediction results are not very stable. When R gets larger, the class imbalance problem becomes more severe and the overall performance drops.

VI. CONCLUSIONS

In this paper, we propose a feature-based method to predict user participation behavior in social networking sites, using only the information of thread content and publisher. To solve this prediction problem, we utilize four types of features, namely thread features, content similarity, user behavior and social features. We further define the "class imbalance ratio" R to alleviate the class imbalance problem, and use an ensemble of LR models for prediction. We conduct an empirical study on Douban Group dataset and the preliminary experimental results demonstrate the effectiveness of our method. In our future work, we plan to explore a wider range of features and analyze which feature contributes most to the prediction task, and further refine our model to facilitate applications in social media analytics [7] and social computing [8].

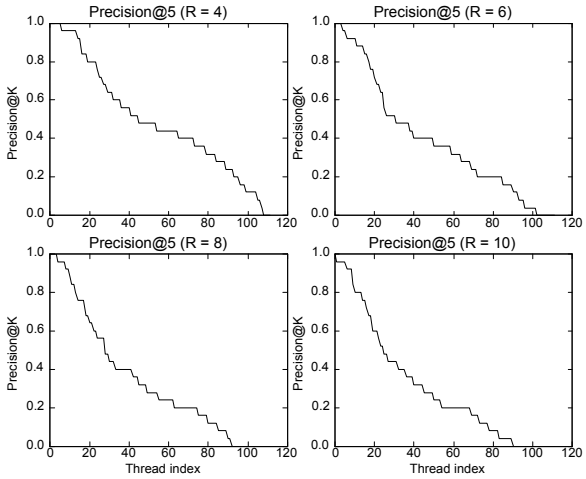

Fig. 1 Precision@5

VII. ACKONWLEDGEMENTS

This work is supported in part by NNSFC grants #61175040, #71025001, #91024030 and #70890084, and MOH grants #2013ZX10004218 and #2012ZX10004801.

REFERENCES

[1] Tsagkias, M., Weerkamp, W., and Rijke, M.d. "Predicting the Volume of Comments on Online News Stories". Proceedings of CIKM 2009.

[2] Yano, T., and Smith, N.A. "What's Worthy of Comment? Content and Comment Volume in Political Blogs". Proceedings of ICWSM 2010.

[3] Hong, L., Dan, O., and Davison, B.D. "Predicting Popular Messages in Twitter". Proceedings of WWW 2011.

[4] Yano, T., Cohen, W.W., and Smith, N.A. "Predicting Response to Political Blog Posts with Topic Models". Proceedings of HLT-NAACL 2009.

[5] Tang, X., Yang, C.C., and Zhang, M. "Who will be Parcipating Next? Predicting the Parcipation of Dark Web Community". Proceedings of ISI-KDD 2012.

[6] Blei, D.M., Ng, A.Y., and Jordan, M.I. "Latent Dirichlet Allocation", Journal of Machine Learning Research, 3: 993-1022, 2003.

[7] Zeng, D., Chen, H., Lusch, R. and Li, S. H. "Social media analytics and intelligence". IEEE Intelligent Systems, 25(6):13-16, 2010.

[8] Wang, F. Y., Carley, K. M., Zeng, D. and Mao, W. "Social computing: From social informatics to social intelligence". IEEE Intelligent Systems, 22(2):79-83, 2007.