PROCEEDINGS OF SCIENCE

Fake Comments Detection Method Based on Integrated Features

Kao Zhang[1]
National Digital Switching System Engineering & Technology R&D Center
Zhengzhou, 450002, China
E-mail: zhangkao@hotmail.com

Ruifei Cui
National Digital Switching System Engineering & Technology R&D Center
Zhengzhou, 450002, China

In view of the problem that existing fake comments detection methods did not take full advantage of dynamic information contained in the history of user behavior, in the paper, we firstly mine the dynamic features from the information of user dynamic behavior by using the time series analysis model. Secondly, we integrate the dynamic features and the user-level static features to find the suspicious users, then the user suspicious probability is propagated to the user comments. Finally, we get the fake comment classification features by fusing comment suspicious probability and the comment-level static features, and then use the PU-Learning classification strategy to accomplish the detection of the fake comments. Experimental results show that, the method we proposed can effectively improves the accuracy of fake comments detection system.

CENet2015
12-13 September 2015
Shanghai, China

[1] Speaker

1. Introduction

Recently, electronic commerce has quickly become the mainstream of global consumer market owing to its advantages of the globalization of markets, continuous trading, low cost and so on. People have begun to become more and more dependent on Internet resources to provide decision supports for themselves. Therefore there are enormous commercial interests in the comments of goods in Internet electronic commerce.

Drived by the interests, there are more and more fake comments beginning to emerge. Professor Liu pointed out that nearly one third of the comments on Amazon were fake [1]. Michael Luca indicated that the percent of fake comments in the Yelp was about 16% [2]. Such a large number of false comments have very bad harm. In order to eliminate the harm and purify the environment of the Internet, it is imminent to develop high-performance fake comments detection method.

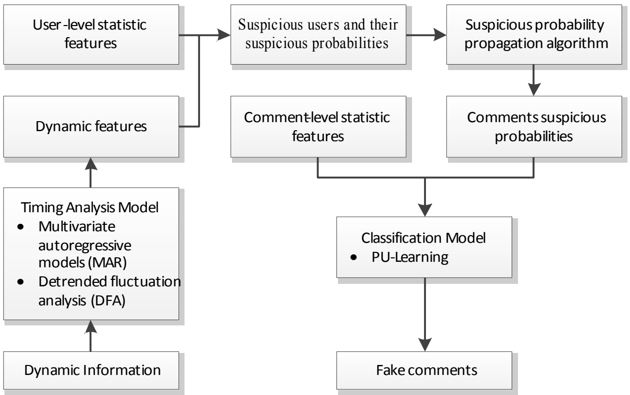
User -level statistic features
Suspicious users and their suspicious probabilities
Suspicious probability propagation algorithm
Dynamic features
Comment-level statistic features
Comments suspicious probabilities
Timing Analysis Model
Multivariate autoregressive models (MAR)
Detrended fluctuation analysis (DFA)
Classification Model
PU-Learning
Dynamic Information
Fake comments

Figure 1: The structure diagram of fake comments detection method based on integrated features

Currently, the supervised learning method has been widely used to detect fake comments. It was first put forward by Jindal N and Bing L. In their mind, duplicate was a major characteristic of fake comments, so duplicate comments were used as fake comments and the rest of comments as non-fake comments in the training data for the logistic regression model. However, in the practical application, it is a nearly impossible task to judge the authenticity of a comment, so the practicability of the method was poor. In order to solve the problem, Hernandez etc. [3] put forward a method of detecting fake comments by using PU-Learning (Positive-Unlabeled Learning) [4-5] strategy and the method greatly improved the recall rate of the detection system. Li etc. made a further improvement of the method and put forward a kind of collective classification algorithms named MHCC (Multi-typed Heterogeneous Collective Classification) by adding users and IP address information, then extended it with the PU-Learning, obtaining a better prediction result [6-7].

From the above analysis, we know that some current studies focus on the comments themselves and make in depth analysis about the difference between fake and true comments to detect fake comments. Despite some achievements, all these studies have a common shortcoming that the features they apply are mostly static features. As a matter of fact, the history of user behavior contains rich dynamic information which can be used to extract dynamic features describing the trend and characteristics of the user behaviors, further integrating these dynamic and static features, the accuracy of detection of fake comments is bounded to be improved. In the paper, we take advantage of this thought to solve the problem of fake comments detection. The fundamental process and structure is shown in Figure1. The main contributions are:

1. Expounding the dynamic information used in the paper and two kinds of time analysis models, and then using the model to extract dynamic features;

2. Integrating dynamic features and static features at the user level to discover suspicious users and then propagating suspicious probability of users to comments published by users;

3. Regarding the suspicious probability of comment as one-dimensional important feature and integrating with static features at the comment level to form fake comments classification feature, then using the PU-Learning classification strategy to realize the detection of fake comments.

2. Dynamic Features Extraction

This section first introduces the dynamic information utilized in the detection of fake comments and then introduces two kinds of timing analysis models which are used to extract dynamic features from the timing information.

2.1 Dynamic Information

This section focuses on the data of user historical behaviors and obtaining timing sequence information from these data. The information used in the paper includes:

1. All comments published by one user are ranked according to the time order and labeled in turn according to the sequence from morning till night (1, 2, 3……). The label is regarded as the independent variable and the score of comments is regarded as the dependent variable, constituting a time series that can be used;
2. Regarding the publication time(pre-processed as the integer from 1 to 24) of all comments of one user as the dependent variable, then obtaining the time sequence of the user active time;
3. Regarding the user host IP as the dependent variable and obtaining the time sequence of the location of the user.

2.2 Timing Analysis Model

2.2.1 Multivariate Autoregressive Models (MVAR)

Every variable of MVAR model is represented as a linear function constructed by the previous values of itself and all other variables. The *p* order and *d* variable MVAR model can be defined by the following equation:

$$\boldsymbol{X}(n)=\sum_{i=1}^{p}\boldsymbol{A}(i)\boldsymbol{X}(n-i)+\boldsymbol{E}(n) \tag{2.1}$$

Where, for the given *d* signals $\boldsymbol{X}(n)=\left(x_1(n),x_2(n),\cdots,x_d(n)\right)^T$, Every signal can be represented by the data collected from different channels, *X*(*n*) is a *d*-dimension column vector representing the values of the multivariable process at the time *n*, *A* (*i*) is the (*d***d*) prediction coefficient matrix, $\boldsymbol{E}(n)=\left(e_1(n),e_2(n),\cdots,e_d(n)\right)^T$ is a prediction error vector.

2.2.2 Detrended fluctuation analysis(DFA)

The detrended fluctuation analysis is an effective method applied to analysis the long range correlation of time and space sequence, and now it has been developed into a kind of statistics methods widely used to the detection of the long range correlation of noise and unstable time series. Through DFA algorithm, we can obtain a scaling exponent and one of its characteristic is that the range of its value reflects the long range correlation between the data of the series. We can use the characteristic of the scaling exponent *α* of DFA to our research.

The coefficients of matrix *A* and scaling exponent *α* describe the dynamics of the system and in this study are used as dynamic features. The reason why *A* and *α* can be used as features for classification is that there is a big difference between the parameters of true comments and that of fake comments.

3. Fake Comments Detection Method Integrated dynamic Features and static Features

Figure 1 shows that the method we are proposing has two integrations of features, 1) Integration of the user dynamic features and static features at the user level; 2) Fusion of comments suspicious probability and static features at the comment level. This section starts with the first step of the integration and how to obtain comment suspicious probability, and then introduces the second step of integration and PU-Learning classification strategy. The proposing method is called as fake comments detection method based on integrated features (IFMFD) because of its fusion characteristic.

3.1 Obtain comments suspicious probability

The obtainment of comments suspicious probability can be divided into two steps: the acquirement of suspicious users and their suspicious probability, the spread of suspicious probability.

3.1.1 The acquirement of suspicious users and their suspicious probability

User-level static features we used in this paper includes the hosts of reviewer, IP, the average score and standard deviation of reviewer; the extraction of dynamic features are described by the section 2.1 and 2.2. The suspicious user obtainment problem can be attributed to a simple problem of supervised learning after extracting the classification features. We employ SVM to accomplish this task due to its high accuracy and good robustness.

3.1.2 The spread of suspicious probability

Li proposed a collective classification algorithm named MHCC (Multi-typed Heterogeneous Collective Classification)[6]. The advantage of MHCC is that it can transfer the user suspicious probability to their comments, which is the intermediate result MHCC algorithm. We employ this characteristic of MHCC to achieve the transformation of suspicious probability from the user to the corresponding comments.

3.2 PU-Learning classification strategy

From the analysis of the introduction, we can know that it is very difficult to construct a large training dataset in a real application scenario, because judging the authenticity of a comment is a nearly impossible task, even for humans. Therefore, the recall rate of traditional supervised learning algorithms (such as Decision Trees and SVM) is generally low. However, PU-Learning is a good method to solve this problem. This section briefly introduces PU-Learning method and how to apply it to the detection of fake comments.

PU-Learning is a partially supervised learning method to solve the two classification problems. The difference between traditional methods and PU-Learning is that the PU-Learning training set is formed by the positive examples and unlabeled examples which contains positive examples and negative examples[4-5]. Fig.2 shows the process of achieving the two-classification through PU-Learning and the traditional methods' process is shown in Fig.3.

As mentioned before, Jindal N and Bing L use the text similarity methods to identify the duplicate comments from the comments set, then treat them as the training set, and finally classify the comments using the logistic regression model[1]. We still adopt the text similarity method to identify duplicate comments in the comment set, and denote these duplicate comments as the positive examples in Fig.2 (same as in Fig.3). The rest of training data are denoted as the unlabeled examples in Fig.2 (denoted as the negative examples in Fig.3). In this way, the recall rate of fake comment detection will be greatly improved.

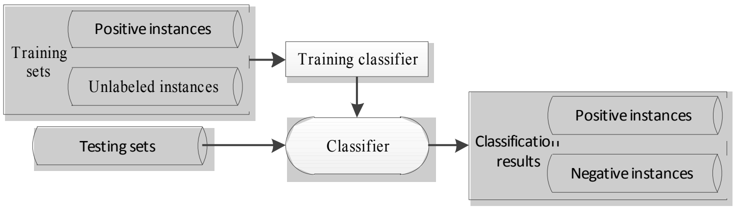

Figure 2: The classification process of PU-Learning

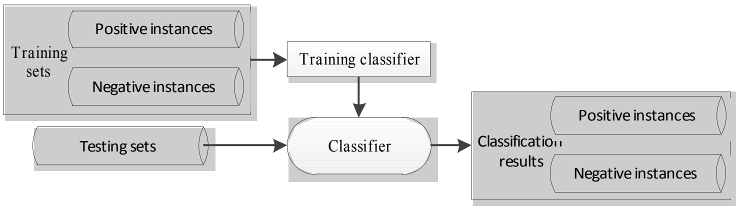

Figure 3: The classification process of traditional classifier

4. Experiment results and analysis

Experimental data is from the Dianping.com and is relatively comprehensive, containing the comment data of 500 restaurants from 2011/11/1 to 2013/11/28 in Shanghai. The primary fields include comment user, IP, comment ID, comment content, comment time, comment score, commented product etc. described the detailed information.

The IFMFD method is a binary classification essentially. The evaluation index of this problem includes recall rate *R*, precision *P*, *F*-Value, where *F* is the geometric average value of *P* and *R*.

4.1 Comparative Experiments

In order to verify the join of dynamic feature can improve the performance of fake comment detection (effectiveness of the method we proposed), the IFMFD method and other two kinds of fake comment detection methods are compared. The methods participated in comparison are listed as the following:

- PU-LEA NB (PU-Learning Naïve Bayes): the classification features used by the method is mainly the comment itself such as the length of comment, comment score, feedback score, the number of useful feedback, comment readability etc.[3]. These features are static characteristics.
- CPU (Collective PU-Learning): This method is an improvement of PU-LEA NB[6]. The difference between these two methods is that CPU regards the IP information and user's own information as the classification features which is still the static characteristics.

methods	True comments			Fake comments		
	P	R	F	P	R	F
PU-LEA NB	0.873	0.738	0.800	0.769	0.884	0.822
CPU	0.886	0.754	0.815	0.780	0.901	0.836
IFMFD	0.885	0.773	0.825	0.782	0.913	0.841

Table 1: The comparative results of fake comments detection methods

The comparisons of the experimental results of three fake comment detection methods are shown in Table1. From the experimental results we can see that: 1) no matter for the fake or true comments, the precision, recall rate and F-Value of IFMFD are higher than PU-LEA NB's and CPU's (for fake comments, F-Value of IFMFD is higher than CPU about 0.5 percent), which means that the addition of dynamic features can effectively improve the accuracy and efficiency of fake comments detection system. 2) The performance of CPU is better than the PU-LEA NB, which also verifies that adding the IP information and user's own information to the classification features can improve detection performance of the system.

4.2 Selection of PU-Learning strategy

Section2.2 briefly introduces the PU-Learning strategy, and describes the reason why the PU-Learning strategy is more suitable for the problem of fake comments detection than the traditional supervised learning methods from the aspect of the principle. This section discusses this conclusion from the aspect of experiment. The comparison of the results of different classification methods at same circumstances of experimental data and features are shown at Table 2. The classifiers participated in this comparison includes Logistic Regression model (LR), support vector machine (SVM) and PU-Learning strategy (PU-LEA).

methods	True comments			Fake comments		
	P	R	F	P	R	F
LR	0.792	0.532	0.636	0.684	0.689	0.686
SVM	0.783	0.543	0.641	0.672	0.702	0.687
PU-LEA	0.885	0.773	0.825	0.782	0.913	0.841

Table 2: The comparison of experimental results of different classification methods

From the data of the table, we can find that, for the problem of fake comments detection, PU-LEA's performance is better than LR and SVM. And we should pay attention to that, the introduction of PU-LEA makes the improvement of the recall rate of the system detection significantly. This is because the fake comments are difficult to get by manually tagging. So the training and test sets are incomplete and inaccurate.

5. Conclusion

In this paper, we first mined the dynamic features from the information of user dynamic behavior by using the time series analysis model. Secondly, we integrated the dynamic features and user-level static features to discover the suspicious users. Then, the user suspicious probability was propagated to the user comments. Finally, we got the fake comment classification features by fusing comment suspicious probability and comment-level static features, and used the PU-Learning classification strategy to accomplish the detection of the fake comments.

References

[1] N. Jindal, B. Liu. *Review spam detection.* Proceedings of the 16th international conference on World Wide Web. ACM, 2007, pp: 1189-1190.

[2] M. Luca, G. Zervas. *Fake it till you make it: Reputation, competition, and Yelp review fraud*. Harvard Business School NOM Unit Working Paper, 2013 (14-006).

[3] D. Hernández, R. Guzmán, M. Móntes, Y. Gomez, P. Rosso. *Using PU-learning to detect deceptive opinion spam*. Proc. of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis. 2013. pp: 38-45.

[4] B. Liu, W. S. Lee, P. S. Yu, X. Li. *Partially supervised classification of text documents*. ICML. 2002, NSF. pp: 387-394.

[5] B. Liu, Y. Dai, X. Li, W. S Li, P. S. Yu. *Building text classifiers using positive and unlabeled examples*. Data Mining, 2003. ICDM 2003. Third IEEE International Conference on. IEEE, 2003, pp: 179-186.

[6] H. Li, Z. Chen, B. Liu, et al. *Spotting Fake Reviews via Collective Positive-Unlabeled Learning*. 2014(899-904).

[7] H. Li, B. L, A. Mukherjee, J. Shao. *Spotting Fake Reviews using Positive-Unlabeled Learning*. Computación y Sistemas, vol.18, no.3, pp: 467-475, 2014.