

# Research on Characteristics Analysis and Classification of Microblog Users

Jiang Xin<sup>1, a</sup>, Li WenMin<sup>2, b</sup>

<sup>1,2</sup>State Key Laboratory of Networking and Switching Technology

Beijing University of Posts and Telecommunications, Beijing 100876, China

<sup>a</sup>csjxsw@bupt.edu.cn, <sup>b</sup>lwm@bupt.edu.cn

**Keywords:** Microblog Users; Feature Detection; Classification Algorithm; C4.5 Decision Tree

**Abstract.** Microblog has become an important part of social media, a large number of users send and thus spread information on this platform. Nowadays, the network environment of microblog is impacted by the presence of anomalies in users seriously. So the research on identifying the types of microblog users is of great significance. Based on the example of microblog, this paper selects some microblog users as research objects and thus analyzes and extracts the features of the selected users. Meanwhile, it uses statistical methods and classification methods in data mining to analyze user data. With the breakthrough point, the classification method C4.5 Decision Tree, this paper has trained the history data to form a classifier to make prognostic classification of new sample, which has realized high accuracy.

## Introduction

In recent years, microblog has become an important part of social network with the rapid development and popularity of microblog in netizens. However, there are a lot of abnormal users in the network created by microblog. This phenomenon has been seriously affecting the normal user experience. Therefore, designing a kind of effective method to identify abnormal users effectively is of practical significance for purifying the Internet environment and improving the user experience.

To some extent, the research on identifying microblog's user type is also the research on the case of user classification. Based on users' socialization and their choice of microblog's text properties, Fabricio et al. held YouTube as the object to study how to identify spam and normal users in video network<sup>[1]</sup>. Gianluca Stringhini et al. studied how to forecast spam in Facebook, MySpace, Twitter through the establishment of user behavior model<sup>[2]</sup>. Liu Kan et al. used Random Forest to do a research aimed at the identification of machine-operated users, which finally achieved good results<sup>[3]</sup>.

The research on microblog users of domestic studies just has got started over the past two years<sup>[4]</sup>, so there are some deficiencies. As for the research on the classification of user type, there are two features: qualitative research methods are widely used while quantitative research methods are seldom used; Most researches concentrate on the identification of certain kind of abnormal users and normal users<sup>[3,5]</sup>, while less researches emphasize on the identification of abnormal users and normal users as a whole.

The latter part of this paper is divided into three parts. The second part is mainly about the theoretical foundation and construction algorithm that the C4.5 Decision Tree involves; The third part is mainly about the extraction of the features of users and data analysis; The fourth part is mainly about using C4.5 Decision Tree to classify user data.

## Related Knowledge

This section describes the relevant concepts of the main classification algorithm, C4.5 Decision Tree. C4.5<sup>[6]</sup> is the algorithm used in the classification of data mining algorithm. It can deal with attributes of discrete type and continuous type with higher accuracy. Info Gain Ratio is the main reason for node split in the process of building C4.5 Decision Tree.

**Info Gain.**

The concept of Info Gain<sup>[6]</sup> comes from information entropy. Entropy means the degree of a system in disorder. The more disordered a system becomes, the higher the entropy gets. Set  $S$  as a set having positive and negative sample. If the target attribute has  $n$  different values, the entropy of  $S$  after classification under  $n$  different conditions is defined as:

$$Entropy(S) = \sum_{i=1}^n -p_i * \log_2 p_i$$

Of which,  $p_i$  is the proportion of  $S$  which belongs to category  $i$ .

Info Gain aims to each attribute. For an attribute  $A$ , the difference between the status under which there is or not  $A$  in the computing system is the amount of information that attribute  $A$  brings about, i.e., Info Gain. The Info Gain of the relative sample set of the attribute  $A$ ,  $S$ , is defined as:

$$Gain(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

Of which,  $Values(A)$  is all possible values for the attribute  $A$ ,  $S_v$  is the subset of attribute  $A$  of  $S$  equal to  $v$ .

**Info Gain Ratio.**

When C4.5 Decision Tree selects features, it will use Info Gain Ratio to overcome the shortcomings when Info Gain prefers to select attributes with more values, and achieves a better result. Info Gain Ratio<sup>[6]</sup> is defined as follows:

$$GainRatio(S, A) = \frac{Gain(S, A)}{SplitInformation(S, A)}$$

Of which,  $Gain(S, A)$  is the Info Gain of attribute  $A$  to set  $S$ .  $SplitInformation(S, A)$  is information separation measure, presenting  $n$  sample subsets from  $S_1$  to  $S_n$ . Use attribute  $A$  to partition  $S$  and then we'll get it. It's defined as:

$$SplitInformation(S, A) = - \sum_{i=1}^n \frac{|S_i|}{|S|} \log_2 \frac{|S_i|}{|S|}$$

**C4.5 Algorithm.**

The process of C4.5 algorithm<sup>[6]</sup> constructing a decision tree is the key step to construct the decision tree classifier. Down from the root node, choose the attribute of maximum Information Gain Ratio as the node to split. Then build the tree one layer after another until all attributes run out at the leaf node. The algorithm is described as follow:

---

C4.5 Algorithm ( $R$ : the no category attribute set,  $C$ : class set,  $S$ : training set)

---

begin

    If  $S$  is null, return a single node equal to failure;

    If  $S$  comprises records of the same categorical attribute value, then return a single node with such value;

    If  $R$  is null, then return a single node. Its value is the categorical attribute value with the highest frequency in the records of  $S$ ;

        for all attributes  $R(R_i)$  do

        if attribute  $R_i$  is continuous attribute, then

        begin

            assign the minimum of  $R_i$  to  $A1$ ;

            assign the maximum of  $R_m$  to  $A_m$ ;

        for  $j$  from 2 to  $m-1$  do  $A_j = A1 + j * (A1 - A_m) / m$ ;

            assign the attribute  $(R_i, S)$  with the maximal Info Gain based on

---

---

```

        {<=Aj, >Aj} to A at Ri;
    end;
    assign the attribute (D, S) with the maximal Info Gain among the attributes of R
    to D;
    assign the value of attribute D to {dj/j=1, 2...m};
    assign the subset of S made up of records of the value dj corresponding to D
    respectively to {sj/j=1, 2...m};
    return a tree and mark its root as D; mark its brunch as d1, d2...dm;
    then construct follow trees respectively:
    C4.5(R-{D}, C, S1), C4.5(R-{D}, C, S2)...C4.5(R-{D}, C, Sm);
end C4.5

```

---

## The Extraction of User Characteristics

One important step of the research on the classification of microblog users is the extraction of user characteristics. This section combines the selection of user characteristics<sup>[2-4]</sup> and introduction of every field in sina API in previous research. Then it selects user characteristics and analyze the extracted features by the means of statistics.

### The Extraction of User Characteristics.

User characteristics extracted from sina API can present the types users belong to roundly. However, not all field information is meaningful. Considering the three important factors, microblog users, the relationship between users and microblog itself, and three concepts, social relations, behavior patterns and the content, we extract 11 user characteristics.

#### 1) Attributes Based on Social Relations

**Friends Count.** There are many differences in the average number of the friends count of normal users and abnormal users. Abnormal users follow lots of other users to achieve the purpose of spreading information.

**Followers Count.** Normal users maintain normal social relations. Their microblogs are so colorful that they attract many people to follow. So abnormal users can't have such many followers.

**Bi Followers Count.** In the process of spam detection, foreign scholars put forward the feature as *Ffratio* in their paper<sup>[2]</sup> to describe the features of users on social relations. The calculation formula is as follow:

$$ffratio = \frac{friends\_count}{bi\_followers\_count}$$

#### 2) Attributes Based on User Behavior Patterns

**Statuses Count.** Abnormal user always release some advertising or useless articles, they can't continue to publish microblog like normal users. So the total number of their microblogs is different.

**Average message sent.** This attribute combines the field *statuses\_count* and the field *created\_at* of user interface. We divide the time (month) from the registration date and the data collection date by the total number of microblogs, and then we will get the average messages sent of users.

**Interval Time.** This attribute gets the time of the latest microblog through the field *status* in user interface. Then it calculates the time difference of data collection. Abnormal users are quite different from normal users in this attribute because of the existence of the *Zombies*. *Zombies* refer to the fake fans in microblogs. These fake fans can be bought by money.

**Forward Rate.** Divide the field *statuses\_count* of user interface by the number of forwarded microblogs and we'll get this attribute. Whether microblogs can be forwarded is decided by the interface field *reposts\_count* of microblog. Among abnormal users, some advertisers produce lots of forwarded information through the forwarding of sales information of online shops on purpose.

### 3)The Attributes Based on The Content of Microblogs

Url ratio. Divide the field statuses\_count of user interface by the number of all microblogs with links and we'll get this attribute. The interface field text of microblogs will decide whether there is a link in a microblog.

Average comments. This attribute presents the mean value of the number of comments. The number can be got through the interface field comments\_count of microblogs. There are interactions between normal users and their friends. On the contrary, abnormal users don't have real friends.

Average length. This attribute presents the mean value of the length of microblogs of user. The length of microblogs can be got through the interface field text of microblog. Since abnormal users put so much ad information in their microblogs, the length of their microblogs is much longer. Normal users won't involve so much information in their microblogs.

#### The Analysis of User Characteristics.

The selected user characteristics should have differentiation in terms of statistical properties. We adopt three commonly used statistical parameters: arithmetic mean, 10% trimmed mean, coefficient of variation<sup>[7]</sup> to distinguish normal users and abnormal users.

This paper get 4137 user samples as data source for research. With the method of manual marking, artificial classification is made for the total sample, including 3613 positive samples and 524 negative samples. Calculating its statistical properties, this paper makes a list for such as shown in table 1.

Table 1 The Attribute Statistics of Experimental data

Statistical Properties User Characteristics		Arithmetic Mean		10% Trimmed Mean		Coefficient of variation	
		Normal	Abnormal	Normal	Abnormal	Normal	Abnormal
Social Relations	Friends Count	527.8	1065.05	420.27	1069.7	1.06	0.57
	Followers Count	6954.12	1145.23	357.37	201.76	17.68	8.46
	Bi Followers Count	190.24	121.22	123.49	43.18	1.55	2.46
	Ffratio	18.05	321.18	7.01	211.52	3.76	1.48
Behavior Patterns	Statuses Count	1432.98	817.4	797.79	344.64	2.21	2.76
	Average Message Sent	5.38	9.83	3.54	4.71	2.14	3.94
	Interval Time	152.49	154.46	87.33	117.9	1.51	1.09
	Forward Rate	0.51	0.61	0.51	0.64	0.61	0.67
Content	Url Ratio	0.21	0.33	0.16	0.29	1.20	1.13
	Average Comments	1.39	0.4	0.71	0.21	4.94	3.62
	Average Length	94.83	108.22	94.7	108.52	0.35	0.3

As is shown, according to different user attributes, normal users and abnormal users have obvious differences in statistical properties. This shows the differentiation of selected users.

### C4.5 Decision Tree Classification Experiment

Before making a C4.5 Decision Tree train,we should do the attribute reduction firstly. Attribute reduction, i.e. feature selection. It refers to removing redundant or unrelated condition attributes from the whole to leave some important condition attributes behind. The reasonable dimension reducing methods mentioned in the literature [8] will make it appear on most classifiers that the result can be improved and near stationary with the increase in the number of features. However, if the number is too large, performance will decrease.

C4.5 Decision Tree classifies on the basis of Info Gain Ratio as a node splitting, this paper adopts the method of Info Gain to deal with attribute reduction. Making use of the Info gain formula mentioned in section 2.1 to calculate the value for Info Gain of 11 user properties, we rank the importance of the attributes.

Based on the ranking, this paper uses different classification algorithm to evaluate and analyze the results of ranking. Then it get the trend based on the ranking that the number of attributes changes

with every evaluation index. This paper compares BayesNet, C4.5 Decision Tree, Logic, SMO, Adaboost, Random Forest, and other classical classification algorithm in data mining<sup>[9]</sup>.

In data mining, Precision, Recall, F-Measure<sup>[10]</sup> are often used to evaluating the result of classification. The figure 1-3 under different algorithms show that based on the ranking of Info Gain, the number of attributes changes with weighted Precision, Recall and F-Measure.

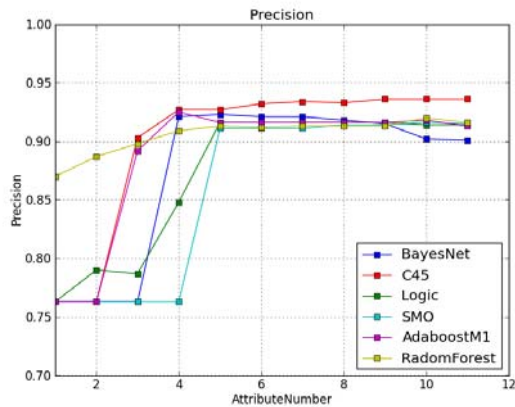


Figure 1 Precision/Attributes

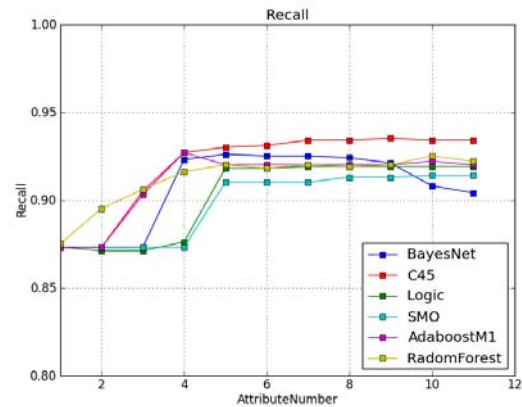


Figure 2 Recall/Attributes

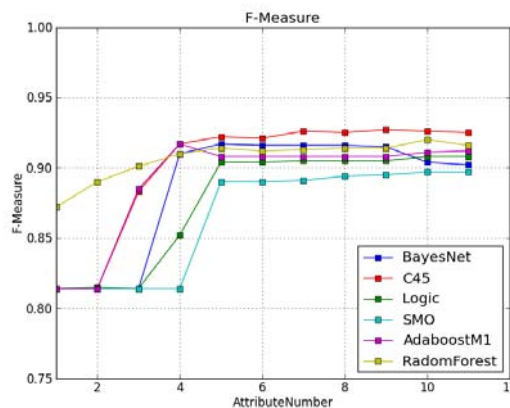


Figure 3 F-Measure/Attributes

As is shown in the figures above, evaluation indexes don't increase with the increasing of the number of attributes. When the number of attributes is greater than a certain value, the curve flattens. This illustrate that, the attribute set after reduction can achieve the result of total attribute set, the 11 attribute can be cut to 5 to replace all attributes.

Combining the figures, C4.5 Decision Tree achieves the best result. So C4.5 Decision Tree algorithm is used to build classification rules for the classification training of user data. The process of constructing a decision tree is just like C4.5 Decision Tree algorithm in 2.3. Select the attribute with the maximal Info Gain Ratio from the continuous values as the split node. Firstly, select the attribute with maximal Info Gain Ratio as the split node in first layer. Then find out the attribute from other properties with the maximal Info Gain Ratio to be the node of next layer. After traversing all attributes, the decision tree forms. Each leaf node in the tree has the function of decision.

This paper use Java to write classification program invoking weka.jar. It works in the form of 10-fold cross-validation. The decision tree contains 9 leaf nodes to decide and the decision-making system can reach the accuracy of 92.68% when predicting user type. It has reached a quite high level in the supervised classification systems.

## Conclusions

This paper takes Sina Microblog as an example to do a research on its users according to the features of microblog users. After obtaining the actual user data, this paper extracts user characteristics. Then it selects the differentiated features to analyze. Last, using method of C4.5

Decision Tree classification helps achieve a good result.

This paper emphasizes on the attribute of samples and the research on the number of attributes and the results of classification. It doesn't consider too much about the influence of the ratio of positive samples to negative samples on the classification results. It will cause the decision-making system tending to identify normal users if positive samples take up too much proportion. Wish researches on these topics can go much further.

## Acknowledgements

This work is supported by NSFC (Grant Nos. 61300181, 61502044), the Fundamental Research Funds for the Central Universities (Grant No. 2015RC23).

## References

- [1] Fabrício Benevenuto, Tiago Rodrigues, Virgílio Almeida, Jussara Almeida and Marcos Goncalves. Detecting Spammers and Content Promoters in Online Video Social Networks [J]. SIGIR'09, July 19–23, 2009.
- [2] Gianluca Stringhini, Christopher Kruegel, Giovanni Vigna. Detecting Spammers on Social Networks[J]. ACSAC '10 Dec. 6-10, 2010.
- [3] Liu Kan, Yuan Yuning, LIU Ping. A Weibo Bot-users Identification Model Based on Random Forest [J], Acta Scientiarum Naturalium Universitatis Pekinensis, 2015, 51(2).
- [4] Peng Xi-xian, Zhu Qing-hua, Liu Xuan. Research on Behavior Characteristics and Classification of Micro-blog Users——Taking“Sina Micro-blog”as an Example [J], Information Science, 2015, 33(1).
- [5] Guo Hao, Lu Yuliang, Wang Yu, Yang Bin. Detection of spam mutual concerns in micro-blogs based on multi-features [J], CHINA SCIENCE PAPER, 2012, 7(7).
- [6] Tom M. Mitchell. [M]. China Machine PRESS, 2003.
- [7] John A. Rice. Mathematical Statistics and Data Analysis[M]. China Machine PRESS, 2007.
- [8] Su JS, Zhang BF, Xu X. Advances in Machine Learning Based Text Categorization [J], Journal of Software September 2006.
- [9] Xindong Wu, Vipin Kumar. The Top the Algorithms in Data Mining [M]. Tsinghua University Press, 2013.
- [10] Yiming Yang. An Evaluation of Statistical Approaches to Text Categorization[J]. Kluwer Academic Publishers, 2000.