http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 13 (7): 1431-1434, 2014
ISSN 1812-5638 / DOI: 10.3923/itj.2014.1431.1434

Microblogging User Feature Analysis based on Boolean Classification Operations

Bing Li, Bingjie Sun, Xuan Wang, Xintong Huang and Xiaoyu Xiu
University of International Business and Economics, Beijing, 100029, China

Abstract: Due to the advance of many social network applications, social group feature analytics are attracting a lot of attention. In the meantime, microblogging, as a kind of social network application, attracts more and more people to use it. With the utilization of bigger and broader crowds over microblogging, surveying massive user features will be an important aspect of exploitation of crowd-sourced data. For better understanding microblogging user group features, in this study, a user classification approach was proposed by means of Boolean operations and it is easily find different microblogging user group features by this approach. In the experiment, some facts were discussed on the exploratory survey to exploit a great deal of microblogging data and how to analyze the features of the different user groups.

Key words: Microblog, user feature, classification, boolean operations

INTRODUCTION

In recent years, the emergence of microblogging website for encouraging crowds to share their information in almost real time across the open space; it is already an important communication and information exchange platform for internet users. However, how to make good use of mass of information brought by the new media is a new challenge.

To study marketing and recommendation based on microblogging, a top priority, it is necessary to study the characteristics of the user groups of the microblogging. Therefore, this study focuses on analysis of microblogging user group characteristics.

As a newly emerged type of social network, Microblogging is increasingly extending its role from communication way into an important platform for sharing real-time information (Hughes and Palen, 2009).

In the first applications area, microblogging user is generally envisioned as social sensors. In the picture, everyone in the microblogging can be viewed as a sensor. Sakaki *et al*. (2010) studied the real-time feature of microblogging user's posts and proposed an event notification system able to monitor posts and deliver notifications promptly. Demirbas *et al*. (2010) proved the precision of using Twitter to deduce information from two case studies.

In the second application area, microblogging user features is used to enhance recommendations. Diakopoulos and Shamma (2010) proposed an approach based on tweet data analysis, to help professionals better understand the dynamics change of people's sentiment. Chen *et al*. (2010) presented an method to rate tweets mainly based on the popularity of embedded URLs. To support tweet ranking, Duan *et al*. (2010) classified tweet features into three categories: content relevance features, twitter specific features and account authority features.

The related work has proved the effectiveness and efficiency of leveraging microblogging to facilitate information dissemination and recommendation. However, according to the literature, the research on microblogging user groups feature analysis is just started. In this study, a novel user group features analytical method was proposed that helps companies to better understand microblogging users crowd characteristics.

MATERIALS AND METHODS

In this study, an efficient method for microblogging users classification based on Boolean algebra was proposed. K-means cluster approach was also applied to do further analysis on important individual user groups.

The data sample in the research comes from weibo-the largest Chinese microblogging website. By means of firefly (Fig. 1), independently developed software of microblogging data crawling by the authors, the researchers randomly collect data from weibo users. Considering the influence of "Zombie Fans" on analysis, data-filtering prerequisites was set in pre-processing to clear away "Zombie Fans" as follows: Fans Number $>\delta_1$, microblogging number $>\delta_2$, Friend Number $>\delta_3$ (δ_1, δ_2, δ_3 are slightly larger than zero).

After the preprocessing, microblogging users matrix was established and do the classification based on Boolean algebra as follows:

Corresponding Author: Bing Li, University of International Business and Economics, Beijing, 100029, China

Fig. 1: Microblogging crawling software, firefly

- Establish n-tuple ui = $\{a_{i1},\cdots,a_{ij},\cdots, a_{in}\}$ to denote all features j of user i, i = 1, 2,···, m; j = 1, 2,···, n. And the n-tuple $\lambda_i = \{\lambda_{i1},\cdots, \lambda_{ij},\cdots, \lambda_n\}$ is established to describe the corresponding attribute value of above-mentioned specific features of user i
- Let θ_j is the median of the set $\lambda_i = \{\lambda_{i1},\cdots, \lambda_{ij},\cdots, \lambda_n\}$ i, i = 1, 2,···, m; j = 1, 2,···,n. Define a new variable 0 η_{ij}:

$$\eta_{ij} = \begin{cases} 1, & \lambda_{ij} > \theta_j \\ 0, & \lambda_{ij} \le \theta_j \end{cases} \quad (1)$$

  Through Eq. 1, to map λ_{ij} to η_{ij}. And then, to use 0-1 variable to represent the features use i, $u_i = \{\eta_{i1},\cdots, \eta_{ij},\cdots, \eta_{in}\}$, η_{ij} is 0 or 1, i = 1, 2,···, m
- The microblogging users features matrix is now a 0-1 matrix: $U = \{u_1,\cdots, u_i,\cdots, u_m\}^T$, i = 1, 2,·····, m
- A Boolean operation classification algorithm (BOC) was designed for classifying microblogging users into different classes. Firstly, a set was defined as follows:
- $S = \{B_f|B_f = \{b_{f1},\cdots, b_{fj}, \cdots, b_{fn}\}\}$ And, f = 1, 2,···,g,···, h,···,p; j = 1, 2,···,n; And b_{fj} is 0 or 1, the number of b_f is 2^p; To any B_g and B_h, $\exists b_{gj}$, $b_{hj} \Rightarrow b_{gj} \wedge b_{hj}$ are not zero at the same time

Secondly, based on the definition and explanation above, Pseudo code algorithm of BOC based on Boolean operation is as follows (Fig. 2).

On the basis of the BOC classification, K-means method was used according to its specific item attribution by each class of users respectively and then extract the clusters with relatively larger amount of elements as the representatives for typical analysis, aiming to find out the typical characteristics of each class of users, which would provide better service for understanding microblogging users at the meantime.

```
Procedure BOC (u_i, B_f) {
 for (f = 1 to f = p)
  for (i = 1 to i = m)
  {
    if (u_1-B_f == 0) // if the difference is zero
    then (u_1's class lable is B_f);
  }
   next
  next
for (f = 1 to f = p)
output: user u_1 ⇒ λ_i, λ_1 = {λ_i1, λ_ij, λ_in},
   and add λ_i to class lable B_f;
  next
}
```

Fig. 2: Pseudo code of BOC

Basic algorithm of K-means is as follows

Choosing K points as the initial centroid
{
Repeat:
Assign each dot to the nearest centroid to form K clusters based on Euclid distance
Recalculate the centroid of each cluster
Until: the variation extent of the centroid is below the given number
}

Through this method, it is easy to find the typical user cluster in one specific user group and then, sampling and descriptive statistical analysis can done from this

typical user cluster and finally summarize the typical characteristics of these user groups.

EXPERIMENTS

Experiments study and analysis: Basing on the above theoretical model, an experimental analysis can be done on the features of weibo users. weibo is the largest Chinese microblogging website, it owns nearly 500 million users. By analyzing main features of weibo users, this study proposes a classification idea based on some users features as follow: friends number, fans number, posts number and their total online time. With the weibo API, information was collected from 42079 randomly selected users including information listed above, as well as some additional user information (e.g., verified user or non-verified user).

In addition, taking into account the influence of 'zombie fans' on integral mean by setting up the following filter conditions when calculating the integral mean: number of fans $>\delta_1$, microblog number $>\delta_2$, Friend Number $>\delta_3$. After data preprocessing, 34050 user records were met the filter conditions. Table 1 is an example; it shows the mean, mode and median of these four variables of some users meeting the filter conditions (only non-verified users included, part of the total population).

The Boolean operation classification algorithm mentioned in chapter was adopted, after classification, the proportions of each user class were shown in the following Fig. 3.

Table 1: Statistical characteristic of qualified non-verified users

Analysis	No. of friends	No. of fans	No. of posts	Post frequency (per day)	Total time
Mean	445.20	9649.61	524.01	2.4200	2.288
Median	253.00	190.00	94.00	0.7530	1.867
Mode	2000	2	1	1.9250	1.617

Table 2: Clustering analysis example of a specific type of non-verified users

	Cluster				
Parameters	1	2	3	4	5
No. of friends	680	623	860	661	1135
No. of fans	2067604	3055894	8717	6141616	836145
No. of micro-blogs	9945	14627	1648	13528	5044
Hours of use	6.132	6.132	3.983	6.990	4.682
No. of users	8	7	4455	1	67

Circle represents the cluster with the largest number of element as an example of analysis

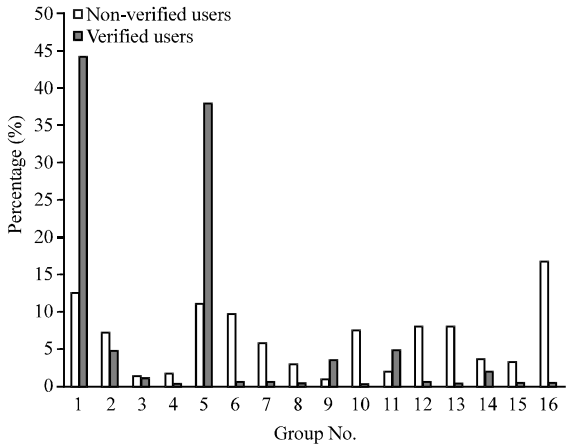

Fig. 3: Proportion of each user category

Deep analysis of the classification result: By taking the K-means method to cluster each type of users, aiming to find the most representative and typical group cluster of each type of users (the cluster which has the largest amount of element in one class) and making analysis of these representative group cluster.

Each type of users was clustered into five clusters and choose the clusters with the largest number of element as a example of analysis, one of the example of the results can be seen as in Table 2.

Use this clustering method to analyze all 16 types of users group. Then sample among those clusters which contain most users of each type, followed by artificial check towards the sample.

According to the analytical result, it could be find out that there are obvious differences between non-verified and verified users:

- Non-verified users include organization users and individual users. Organization non-verified users mainly write micro-blog and gather information. Their micro-blogs are often related to a specific topic. Those who gather information are mainly small organizations and they have the demand for information from other users in order to do research on certain industry. Individual non-verified users are mainly normal users. These users mainly follow their interested users to get information they want, write micro-blog about their daily life and use it as a tool to communicate with friends in real world
- The verified users are rather centered in term of user type, mainly focused on (0, 1, 1, 1) type and (1, 1, 1, 1) type. These two types of users share the same feature of having a huge quantity of fans. They are usually well-known organizations or famous persons in different field and thus be invited to join weibo. These users already have a great influence in real world, so they can easily attract a huge number of followers. They use it as a platform to release information to the public. They are the most social active users on weibo and they play the role as the central nodes in microblogging social network

CONCLUSION

This study introduced a novel approach to analyze microblogging users group features. Especially, this study also do experimental study to prove the validity of the approach based on a real dataset of weibo users. In the future work, much deeper user group features analysis approach will be explored and tested for big data.

ACKNOWLEDGMENT

This study is supported by a grant from Humanities and Social Science Fund of Chinese Ministry of Education (No. 11YJA870017).

REFERENCES

Chen, J., R. Nairn, L. Nelson, M. Bernstein and E. Chi, 2010. Short and tweet: Experiments on recommending content from information streams. Proceedings of the 28th International Conference on Human Factors in Computing Systems, April 10-15, 2010, Atlanta, Georgia, USA., pp: 1185-1194.

Demirbas, M., M.A. Bayir, C.G. Akcora, Y.S. Yilmaz and H. Ferhatosmanoglu, 2010. Crowd-sourced sensing and collaboration using twitter. Proceedings of the IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks, June 14-17, 2010, Montrreal, QC., pp: 1-9.

Diakopoulos, N.A. and D.A. Shamma, 2010. Characterizing debate performance via aggregated twitter sentiment. Proceedings of the 28th International Conference on Human Factors in Computing Systems, April 10-15, 2010, Atlanta, GA., USA., pp: 1195-1198.

Duan, Y., L. Jiang, T. Qin, M. Zhou and H.Y. Shum, 2010. An empirical study on learning to rank of tweets. Proceedings of the 23rd International Conference on Computational Linguistics, August, 2010, Beijing, China, pp: 295-303.

Hughes, A.L. and L. Palen, 2009. Twitter adoption and use in mass convergence and emergency events. Int. J. Emergency Manage., 6: 248-260.

Sakaki, T., M. Okazaki and Y. Matsuo, 2010. Earthquake shakes Twitter users: Real-time event detection by social sensors. Proceedings of the 19th International Conference on World Wide Web, April 26-30, 2010, Raleigh NC., USA., pp: 851-860.