Journal of Computers Vol. 29 No. 2, 2018, pp. 186-192
doi:10.3966/199115992018042902018

A Hierarchical Partition Method for User Influence on Micro-blog Users

Yan-li Liu[1*], Xiao-jun Wen[1]

[1] School of Computer Engineering, Shenzhen Polytechnic, Shenzhen, Guangdong 518055, China
teach_ie@163.com

Received 24 July 2017; Revised 19 September 2017; Accepted 19 October 2017

Abstract. The user hierarchical partition is an important part of the public opinion analysis. To solve the problem of incomplete user influential characteristics, multi-dimensional information mining is used to analyze the factors of user influence hierarchy by integration of user subject similarity, emotional tendency, community influence and other characteristics. A hierarchical partition method of user influence based on fusion feature similarity is proposed. Experiments from real micro blog data show the effectiveness of the proposed method.

Keywords: feature fusion, hierarchical partition, micro blog analysis, user influence

1 Introduction

Micro blog is an important platform for web users to describe their interests, express their attitudes, share and discuss with others by sending short messages. Users can integrate various information channels through micro blog (including instant messages, mobile phones, Emails, web pages, etc.) and release texts, pictures, videos, audios and other forms of information. Therefore, micro blog analysis plays an important role in the field of public opinion analysis. Since the user hierarchical partition has an important role in key user digging, the user hierarchical analysis has a high guiding position in the micro blog public opinion area [1-2].

The main works of micro blog user influence hierarchical analysis include: researching the basic behavior features and relationship features of micro bloggers, analyzing the relevant variables of micro blogger influence, and establishing a model associated with the number of fans and concerns, the number of blogs, and other factors. At the same time, when we track the topical data that micro bloggers exchange in a certain period, we can find that attentions, comments, forwarding and citations etcetera can form social relationship networks while users exchange information. In the relationship networks, user communities can be discovered, and user hierarchies can be partitioned.

Qiao et al. [3] found the core members in the network by using the distribution model with email communication behaviors to dig user email data and found the personality characteristics of users. Nascimento et al. [4] considered the number of papers and number of their citations as an important indicator of the author's influence in the academic network. Agarwal et al. [5] measured the individual influence by considering factors such as user activity, individual expression ability, reputation and novelty together in the micro blog data environment. Hui et al. [6] considered the user emotional tendency, regarded the credibility of users as an important factor in the analysis of user influence. Cai [7], Lin et al. [8] analyzed user influence in different areas. Dietz et al. [9] analyzed the user influence with blog texts based on the LDA model. Literature et al. respectively, to consider the number of fans, forwarding, posts, replies and son on to calculate the individual user influence of community [10-11].

Ding et al. [12] divided influential individuals into "multi-topic level influential individuals" and "single topic level influential individuals" by their influential attributes in the multi-relationship network. They proposed an influential individual discovery method to analyze the user influence from the subject level. Dou et al. [13] established a subject level user influential measurement model associated with the characteristics of user attributes, user behaviors, and information disseminations and so on.

* Corresponding Author

All the above-mentioned methods, which partitioned the user influence hierarchies by the attribute and behavioral characteristics of users, are not fully completed. Thus, we not only concern the characteristics of user attribute and user behavior, we also integrate the user subject similarities, emotional tendencies, community influences and other characteristics. With the integration, we can dig the user information in many dimensions, and analyze the factors affecting the user influence hierarchical partition.

2 The Model of Hierarchical Partition

In this paper, we evaluate the spreading and pushing effect of user influence in consideration of the static properties [14], emotional states [15], and dynamic behaviors [16] of the micro blog users. We analyze the user influence from three aspects: the subject model [17], emotional tendency [18], and community influence. Important individual users are extracted; and key users and user groups are partitioned into hierarchies.

2.1 Subject Influences

Micro blog texts are the main data of user speeches, which include the user emotional status, opinions to the events, attention objects and some other information such as areas. The information characteristics of contexts are usually described by the subject model. The subject model can reflect the characteristics of user opinions and their blog topics, which is an important feature in micro blog analysis. Due to the special nature of micro blog (the length of the text is limited to 140 words or less), the LDA subject model cannot be used directly. In this paper, the LDA subject model is applied on user blog on the foundation of the short text clustering; and calculating the subject similarity is considered an important factor of the user influence hierarchical partition.

The LDA method can generate a document with multiple topics. The model uses the following method to generate the document.

Select the parameter $\theta \sim p(\theta)$;

For each word in the N words w_n;

Select a subject $z_n \sim p(z \mid \theta)$;

Select a word $w_n \sim p(w \mid z)$;

θ is a subject vector, each column of the vector represents the appearance probability of each subject. The vector is a nonnegative normalized vector. $p(\theta)$ is the distribution of θ, especially Dirichlet distribution, a distribution of distribution. N represents the number of words in the generated document. w_n represents the nth word in the generated document. z_n represents the selected subject. $p(z \mid \theta)$ represents the probability distributions of the subject z when θ is given, which is the value of θ, θ: $p(z = i \mid \theta) = \theta_i$. $p(w \mid z)$ represents the distribution of w with the given z, which can be seen as a $k \times V$ matrix. K is the number of subjects, V is the number of words. Each row in the matrix represents the probability distribution of the words corresponding to the subject, which is the probability of each word contained in subject z. With this probability distribution, each word is generated by a certain probability.

With this method, first we select a subject vector θ and determine the selection probability of each selected topic. Then, when each word is generated, we select a subject z from the subject distribution vector θ; and we generate a word according to the probability distribution of the word z at the same time. The model is as Fig. 1.

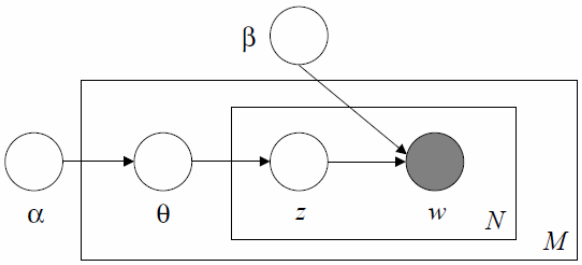


Fig. 1. LDA subject model

From the above Fig. 1, the joint probability of LDA is:

$$p(\theta, z, w \mid \alpha, \beta) = p(\theta \mid \alpha) \prod_{n=1}^{N} p(z_n \mid \theta) p(w_n \mid z_n, \beta) \quad \textbf{(1)}$$

The LDA model is used mainly to train the two control parameters α and β from a given input corpus. When the two control parameters are studied out, the model is ready for generating documents. Through subject modeling of micro blog texts, we can get the subject distribution of each user. With the subject distribution, we can calculate the subject similarity of user blog and hot topics. Analysis of the association between user blog subjects and current hot topics can reflects the subject influence of user blog.

Assume vectors x_i and y_j representing texts *x* and *y* respectively. The text cosine similarity is:

$$T = \cos\_sim = \frac{\sum_{i=1}^{n} x_i \times y_j}{\sqrt{\sum_{i=1}^{n} x_i} \times \sqrt{\sum_{j=1}^{n} y_j}} \quad \textbf{(2)}$$

The processing flow of calculating the blog subject similarity is:

1. Extract the subject words of texts *x* and *y*;
2. For each text, take out a number of subject words, merge them into a set, and then calculate the frequencies of words of the set;
3. Generate the word frequency vectors of x_i and y_j;
4. Calculate the cosine similarity T of the two vectors. The bigger value represents the higher similarity.

2.2 Emotional Tendencies

The textual emotional analysis is that user emotional tendencies can be judged by analyzing and digging the subjective information in the texts, such as expressing emotions, opinions and standpoints. By analyzing the emotional tendencies of the comments from the reviewers, the emotional tendencies of the reviewers can be estimated, which is a key step in the analysis of the social public opinion. At present, the textual emotional tendency analysis is widely used in information retrieval, information filtering, emotion recognition and other fields.

Emotional polarity is an important factor for analyzing the user personal characteristics. In this paper, by analyzing emotional polarity of micro blog with the help of several Chinese emotional dictionaries, we can estimate the emotional attitudes of users to events of micro blog.

In this paper, machine learning and Naive Bayesian model are used to classify user emotions. Suppose *m* is the number of the classes $C_1, ..., C_m$. For a data sample *X*, Taxonomy predicts that *X* belongs to a class C_i, if and only if:

$$P(C_i \mid X) > P(C_j \mid X), 1 \le j \le m, i \ne j \quad \textbf{(3)}$$

According to the Bayesian principle：

$$P(C_i \mid X) = \frac{P(X \mid C_i) P(C_i)}{P(X)} \quad \textbf{(4)}$$

Because the above-mentioned denominator $P(X)$ of the formula is a constant for all classes, the denominator is not considered. During calculation, we only need to maximize $P(X \mid C_i) P(C_i)$ and calculate $P(X \mid C_i)$. Because the Naive Bayesian classification assumption condition is independent, that is, the values of the given sample attributes are independent of each other:

$$P(X \mid C_i) = P(x_1, x_2, ..., x_k \mid C_i) = P(x_1 \mid C_i) \times ... \times P(x_k \mid C_i) = \prod_{k=1}^{n} P(x_k \mid C_i) \quad \textbf{(5)}$$

Then

$$y = f(x) = \arg\max P(C_i) \prod_{k=1}^{n} P(x_k \mid C_i) \quad \textbf{(6)}$$

2.3 Community Influences

The user hierarchical partition is to divide users into different levels according to user interests, speeches, user influences and other features. In order to analyze the user level and fully dig influential data of micro blog, the community influence factor of user personal influences is proposed:

$$Q = \frac{\left(n_z + n_p\right) \bullet n_f}{n_m} + n_z \log\left(n_f + 1\right) \quad \textbf{(7)}$$

In the above formula, n_z represents the number of forwarding; n_p represents the number of comments, n_f represents the number of fans, and n_m represents the number of states. The first part of the factors indicates the influence of forwarding and comments of user blog; the second part of the factors indicates the influence of the inherent community network. With the proposed integrated factor, we can describe the community influence of a user in a period, and the factor can be a basis feature of a user.

2.4 User Influences

The proposed user influential hierarchical partition is to classify users into their belonging levels mainly with the two aspects: the community influence factor and the subject influences. The value of the user influence can be calculated with the following formula based on emotional polarity:

$$J_{\mathrm{i}} = \eta\left(T_i + Q_i\right) \quad \textbf{(8)}$$

$$\eta = \begin{cases} 1, & \textit{Emotional polarity is positive} \\ -1, & \textit{Emotional polarity is negative} \end{cases} \quad \textbf{(9)}$$

In the above formula, T_i represents the subjective influence of user *i*; and Q_i represents the value of the user community influence of user *i*.

3 Experimental Analysis

3.1 Experimental Data

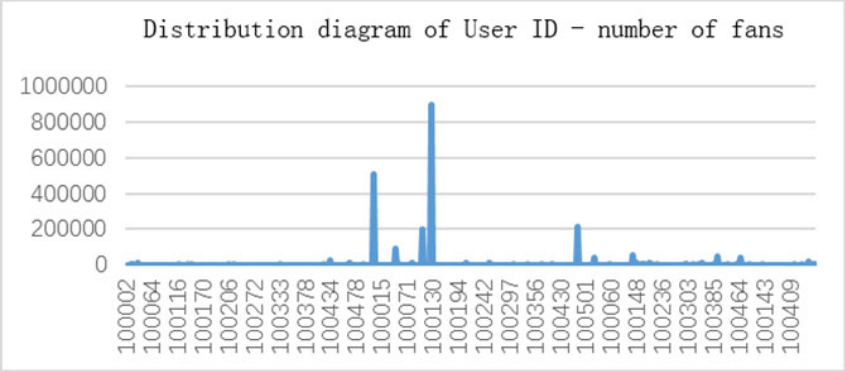
The experimental data contains 5000 micro blog texts from 513 users crawled on October 2014 to June 2015, and each user posted nearly 10 micro blog texts in average. Each blog text contains the basic attributes of users (number of fans, number of states, number of forwarding and comments etc.). The blog texts are shown in Table 1. After analyzing texts of some users, the numbers of user blog and fans are shown in Fig. 2.

3.2 The Analysis of User Influence

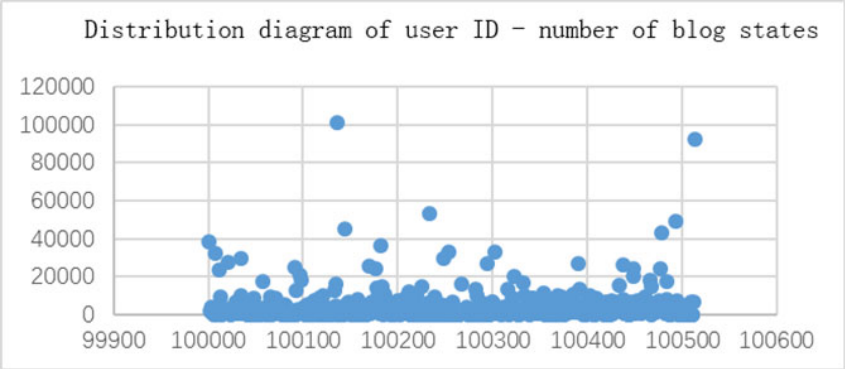
The user influential factor indicates the activity of a user in micro blog community. A factor with greater value means the blog of the user can be spread wider and faster. In this paper, users are classified into 5 levels according to their influential factors. In the 5 levels, 0 is the highest level, indicating the greatest influence on users. The partition rules are shown in the following Table 2.

Table 1. Data instance

ID	Number of forwarding	Number of comments	issuing time	gender	Number of fans	Number of states	Text(original text)
100001	2	0	20150119 21:13:21	m	1889	38143	100 Yuan is enough [sad] //@ TengjiTuoye: 500 @blacksauce_weightlossmarch: more than 200 yuan, with bank cards// @ whipinwhitewater: the amount of money is not important, the bank card with money inside is most important ... @ Japanzerodistance: 100 @ heavytasteofyoungwomen: at least one hundred Yuan, maybe not enough for calling a taxi in Shanghai



(a) Distribution diagram of User ID - number of fans



(b) Distribution diagram of user ID - number of blog states

Fig. 2. experimental process

Table 2. Level partition rules

Level-0	Level-1	Level-2	Level-3	Level-4
Inf >=100	100>Inf >=10	10>Inf >=3	3>Inf >=1.5	1.5>Inf

According to the user hierarchical partition method, the influential level of each user can be estimated, as shown in Fig. 3 and Fig. 4.

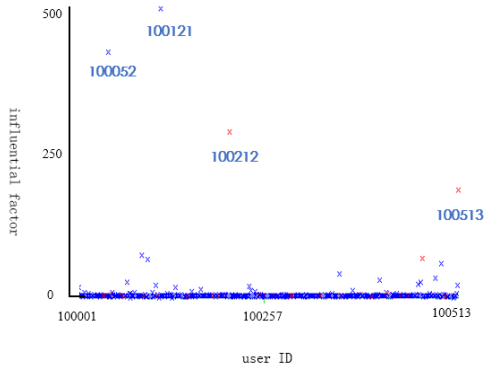
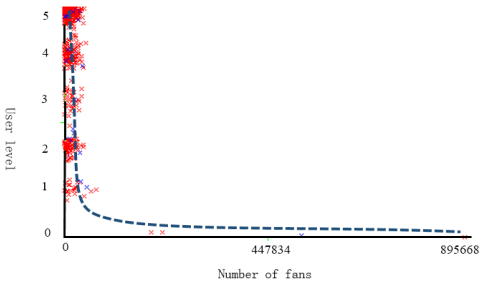
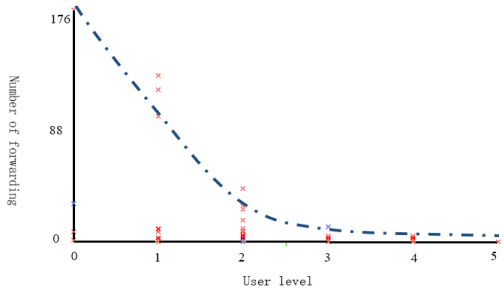


Fig. 3. Experimental process



(a) Figure The number of fans - User level



(b) The user level – number of forwarding

Fig. 4. The distribution of the influential factors in different user levels

In Fig. 3, blue means the emotional polarity is negative, while red means the emotional polarity is positive. From the diagram, it is obvious that there is the largest quantity of users who have the least user influence (5 level), and the least quantity of users who have the greatest user influence (level 0). The distribution falls in the community rules. We can also find that the users who have the greatest influences are User No. 100121, 100052, 100212 and 100513. The emotional polarities of User No. 100121 and 100052 are negative, which present that they often spread the social negative emotions. The emotional polarities of User No. 100212 and 100513 are positive, which present they often spread the social positive emotions.

At last, we verify the distribution of the influential factors in different user levels. In Fig. 4, with the increase of the user influence, the number of fans increases dramatically, and the number of user blog forwarding is increasing continually too. It shows that the numbers of fans have important effects on the user influences. With the greater user influences, the possibilities and numbers of blog forwarding increase relatively.

4 Conclusions

In this paper, we analyze the real micro blog data, and propose a user hierarchical partition method based on emotional subject influence. We integrate three aspects: the user community network attributes, the blog emotional polarities, and the user subject heat together to analyze the user influence. Experiments show that the proposed method has high practical significance, and it is reasonable and reliable to evaluate the user influence.

However, the paper has some deficiencies, such as the hierarchical partition method of user influence is not comprehensive. Our next step is to design a more complex and reasonable user-level partitioning framework.

References

[1] Z. Xiong, The public opinion analysis of micro blog based on web text extraction, [dissertation] Xi'an: Xi'an Technology University, 2013.

[2] J. Li, Research on micro blog key users and user community network mining, [dissertation] Guangzhou: South China University of Technology, 2015.

[3] S. Qiao, C. Tang, J. Peng, W. Liu, F. Wen, Q. Jiangtao, Digging the core of criminal networks based on personality features in a simulated mail analysis system, Chinese Journal of Computer Science 31(10)(2008) 1795-1803.

[4] M.A. Nascimento, J. Sander, J. Pound, Analysis of SIGMOD's co-authorship graph, SIGMOD Record 32(3)(2003) 8-10.

[5] N. Agarwal, H. Liu, L. Tang, P.S. Yu, Identifying the influential bloggers in a community, in: Proc. the 1st ACM Int Conf on Web Search and Data Mining, 2008.

[6] P. Hui, M. Gregory, Quantifying sentiment and influence in blogspaces, in: Proc. the 1st Workshop on Social Media Analytics, 2010.

[7] Y. Cai, Y. Chen, Mining influential bloggers: from general to domain specific, in: Proc. the 13th Int Conf on Knowledge-Based and Intelligent Information & Engineering Systems, 2009.

[8] C.X. Lin, B. Zhao, Q. Mei, J. Han, A statistical model for popular event tracking in social communities, in: Prco. the 16th Int Conf on Knowledge Discovery and Data Mining, 2010.

[9] L. Dietz, S. Bickel, T. Scheffer, Unsupervised prediction of citation influences, in: Proc. the 24th Int Conf on Machine Learning, 2007.

[10] A. Pal, S. Counts, Identifying topical authorities in microblogs, in: Proc. the 4th ACM Int Conf on Web Search and Data Mining, 2011.

[11] M. Cha, H. Haddadi, F. Benevenuto, K.P. Gummadi, Measuring user influence in twitter: the million follower fallacy, in: Proc. the 4th Int AAAI Conf on Weblogs and Social Media, 2010.

[12] Z. Ding, B. Zhou, Y. Jia, L. Zhang,Topical influence analysis based on the multi-relational network in microblogs, Computer Research and Development 50(10)(2013) 2155-2175.

[13] F. Dou, Research on topic level user influences in micro blog, [dissertation] Xi'an: Xi'an University of Technology, 2014.

[14] W. Yu, W. Tao, J. Xu, et al. User information extraction method based on domain ontologies on micro blog, Journal of Yangtze University (Natural Science Edition) 12(10)(2015) 36-40.

[15] L. Xie, M. Zhou, M. Sun, Multi-strategy emotional analysis and feature extraction based on hierarchical structure on Chinese micro blog, Chinese Journal of Information 26(1)(2012) 73-83.

[16] L. Meng, Research on the behaviors of user groups on micro blog, [dissertation] Wuhan: Wuhan University of Technology, 2012.

[17] H. Liu, W. Li, Y. Zhang, Microblog Topic Detection Based on LDA Model and Multi-level Clustering, Computer Technology and Development 6(2016) 25-30.

[18] X. Zhu, Z. Liu, W. Jin, T. Liu, C. Liu, Y. Chai, The hierarchical structure of micro blog emotion classification based on feature fusion, Telecommunication Science 32(7)(2016) 106-114.