International Journal of Hybrid Information Technology
Vol.9, No.4 (2016), pp. 447-460
http://dx.doi.org/10.14257/ijhit.2016.9.4.39

Microblog Users Roles in Topic Diffusion Based On Individual Attribute Features

QiuLi Qin[1], Xing Yang[2*] and Hua Gu[3]
Laboratory of Logistics Technology and Management
School of Economics and Management
Beijing Jiaotong University
[1,2,3]No.3 Shang Yuan Cun, Hai Dian District
Beijing, China, 100044
[1]qlqin@bjtu.edu.cn; [2]13120642@bjtu.edu.cn;[3]guh@mail.bsti.ac.cn

Abstract

With the rapid development of Social Network, Micro-blog has becoming an important way for people communicating with each other and getting shared resource, so the research of Social Network especially Twitter and Micro-blog is crucial. The previous studies most focused on users' characteristic, topical issues, information broadcasting mechanism and so on. But there have been few studies to combine them together and find more valuable conclusions. In this paper, deep study of users' characteristic and their roles in topic diffusion is proposed. By using correlation analysis and K-Means clustering algorithm, ordinary users can be divided into several different types according to the data characteristics. Discriminate function has proved the rationality of classification in theory. By using PKUVIS, a visual analysis of the topic about the lots of Malaysia airline flight MH370 is applied to trace users' roles in topic diffusion. The empirical analysis show that, Micro-blog users assume different roles in the process of topic broadcasting mechanism based on individual attribute features.

Keywords: *Individual Attribute Features; Topic Diffusion; Visualization*

1. Introduction

In recent years, with the rapid development of Internet and increasing users of Social Network, the way of people's living and working has been deeply influenced. Micro-blog has becoming the mainstay of information distribution and transmission by offering users timely access to information and allowing them to publish their views in concise words at any time in any place. According to statistics, Sina Micro-blog platform already has 503 million registered users until to March 2013.

Different from traditional Social Network, Micro-blog, with the open dissemination of information, wide scope of influence, extensive and timeliness, possesses more interactive users, so it is easy to form a hot topic. As discussion in reality, users on Micro-blog platform have different personality traits. Individual attribute features usually affect their performance in the process of topic transmission. Thus researches synthesizing the factors of users' characteristic, topical issues and information broadcasting mechanism have vital practical significance. Identifying the user types and characteristics is helpful to tracking their roles in topic diffusion. Combined with Micro-blog feature, this papers consider that types of users can be identified referring to their characteristic. Therefore quantitative description of users characteristics and analysis of topic diffusion mechanism are the focus of this paper.

The rest of paper is organized as follows. We discuss the related work on Twitter and Micro-blog in Section 2. In Section 3, we introduce the theoretical basis which will be used in our paper, including correlation analysis, K-Means clustering algorithm and

ISSN: 1738-9968 IJHIT

discriminate analysis. Section 4 shows the analysis of users based on their characteristics and build classification function. Section 5 presents visual analysis of hot topic and Section 6 concludes the paper finally.

2. Related Research

With the rapid development of Micro-blog, researchers have paid extensive attention on it. Twitter as the most popular Micro-blog platform has become the key point in the academic circles studies. Many of the literature conducted research on the characteristic of users characteristic and influence, but the method and views of researches are different.

For example, Kaye D.S weetser extract a large number of Twitter users samples to analysis users' motivation, leadership and the use of social media [1]. Some have built a number of visualizations for analyzing Twitter data from different domains, including general geographical institutional awareness [5], disaster and crisis scenarios [6], and political events. Network structure plays an important role in understanding how information propagate in social networks. Ratkiewicz et al. applied network analysis and visualization for detecting meme diffusion patterns in election-related microblogs [7]. Viegas et al. developed history flow visualization to explore temporal dynamics of Wikipedia [11]. Brandes et al. defined the edit network of Wikipedia authors and proposed methods to better analyze their collaboration and competition relationship [12]. The internal researches on the characteristics of Micro-blog users are not too much, Xiao Qiang et. Draw a conclusion that users can be divided into six classes based on analysis of behavior characteristics using data mining method [2]. Wang xiaoguang, who pays attention to the research of Micro-blog contents, finally find that there are highly positive correlation between the number of fans and contents [3].Yu et al. analyzed temporal characters of Micro-blog usage, and found some differences between Micro-blog and Twitter [9]. Qu et al. analyzed how Chinese people use Micro-blog in response to a major disaster in China [10]. Tang et al. designed LifeCircle, as a summary of the long term life status of Micro-blog users [4].

To sum up, most of previous research work was limited to observe characteristics of user attributes, information content or ways of transmission, but ignoring the influence of information transmission. Our work has two major contributions. Firstly we carry on a thorough analysis of users' characteristic. On the other hand, we establish users classification function combined with the feature of user attributes and information dissemination mechanism. In the analysis, we analyze and visualize Micro-blog events with not only users' data, but also collected the original events data. It enables us to get deeper understanding of Micro-blog users and event transmission.

3. The theoretical study

3.1. Data acquisition and preprocessing

3.1.1. Data acquisition: In this paper, data are from Sina Micro-blog. We use web crawler through open API to get access to the dat. In order to ensure randomness of data, acquisition last for three months, which monitor a certain number on a daily basis including continuous monitoring on certain users. Besides, the monitoring time is scattered relatively. The results include the data of 14374 items in June, 6960 items in July and 5015 items in August in 2013, which contains query time, UID, gender, certification, the number of fans, attention, mutual powder, content, collection, and registration time. Part of data are is shown as follows.

Table 1. The Ordinary User Data Description Statistics

Query time	sex	UID	follows	focus	contents	collection	mutual	certificated	Registered time
2013/6/1 9:00	1	2662662321	307	147	83	0	66	0	2012/3/14
2013/6/1 9:00	1	1746309195	369	319	79	0	271	0	2010/5/24
2013/6/1 9:00	1	1693094457	235	265	901	4	148	0	2010/2/11
2013/6/1 9:00	1	2088022077	473	549	264	3	394	0	2011/4/18
2013/6/1 9:00	1	2127598901	420	256	153	2	220	0	2011/5/10
2013/6/1 9:00	1	3115081651	808	905	244	3	711	0	2012/11/15
2013/6/1 2:53	2	1266321801	47122072	580	6759	111	568	1	2009/8/28
2013/6/1 9:00	2	1777146727	511	708	706	5	240	1	2010/7/17
2013/6/1 9:00	2	2129713552	397	340	1389	99	146	1	2011/5/3
2013/6/1 9:44	1	1912386535	28974	1583	9866	18	1121	1	2011/1/5
2013/6/1 9:45	1	2815663917	654	153	661	0	44	1	2012/9/20
2013/6/1 9:48	1	1959182407	98879	1502	6813	142	993	1	2011/2/9
2013/6/1 9:48	1	1898885525	686262	249	17269	8	158	1	2010/12/22
2013/6/1 9:48	1	1709951635	108346	246	11642	8	114	1	2010/3/13

3.1.2. Data preprocessing: Because of the randomness in the process of data acquisition, some users are repeated, so it is necessary to clean up the data. First we delete 193 pieces of information which was grabbed randomly by crawler software and 7183 pieces of repeat information which share the same UID. There are still 19008 pieces of remaining information.

1) There are a large number of certified celebrities in Sina Micro-blog, who can attract a large number of followers without taking any action. So we decide to analyze authenticated users and ordinary users respectively to avoid this kind of disturbance made by this kind of users. Through data filtering, we find that authenticated users, which are 12316 people, take off about half of the total, on the other hand verified the randomness of the experimental data.

2) There exits some users whose follows are nearly zero. These types of users usually don't keep a watchful eye on others. some of them maybe already have their accounts disabled. The other part is this kind of users who treat Micro-blog as a marketing account. They usually pay a lot attention on others and publish advertisement on the platform. So we define a certain threshold at last and then delete data of users whose fan number are less than ten, Micro-blog number are less than five.

3) In order to conduct a comprehensive analysis of the user data, we use one to replace male and two for female as gender in user attributes. We also use Boolean variable express the user authenticated properties, in which one means authenticated users and zero means ordinary users.

After preliminary data cleaning, we still remaining 6277 normal user information, 12250 authenticated user information as the experimental data for empirical research.

3.1.3. Data Description： Since there are significantly different influences between the ordinary users and authenticated users, to get more information from them and have good understanding of the comprehensive characteristics of the data, we have described the data set of the ordinary users and certificated users respectively through SPSS. The results are as follows:

Table 2. The Ordinary User Data Description Statistics

	N	min	max	average	SD	variance
follows	6277	11	13661251	79049.75	594669.985	3.536E11
attention	6277	0	3000	526.06	537.911	289347.916
contents	6277	6	152476	1962.79	5098.802	2.600E7
collects	6277	0	10583	108.43	551.386	304026.601
mutual fans	6277	0	2935	178.97	339.830	115484.310

Table 3. Authenticated User Data Description Statistics

	N	max	average	SD	variance
follows	12250	54595975	2698335.27	7879392.748	6.208E13
attention	12250	3000	467.21	509.826	259922.136
contents	12250	87287	5200.85	7170.158	5.141E7
collects	12250	139607	231.82	1531.062	2344151.997
mutual fans	12250	2943	318.49	388.756	151130.936

Description statistics of Ordinary users and authenticated users show that the two types of users are totally different on the number of follows, attention, contents and mutual fans. In order to eliminate the effect of celebrities and make the analysis more rigorous, we analyze ordinary users and authenticated users respectively in the next work.

3.2. Correlation analysis

Because of the growing needs of participate in the process of information dissemination for audience on Micro-blog platform, ordinary users are playing a more and more important role in information dissemination process. There are a lot of scholars beginning to research the influence of Micro-blog transmission based on the theory of long tail. Therefore, ordinary users, as the main force on Micro-blog platform are getting more attention for their behavior and characteristics. In all attributes of users, the number of attention and contents are important factors to measure the activity of a user on the platform. Besides, number of follows, which is also the factor users usually concerned most, is significant factor to evaluate a user's influence. The number of mutual attention

can be used to measure the social demand of users. It is useful to get to know the number of contents, attention, follows of a user to grasp the characteristics and understand the role a user plays on the Micro-blog platform.

Table 4. Correlation coefficient of number of follows and attention using number of contents as control variables

control variables			follows	attention
contents	follows	correlation	1.000	-.052
		significance	.	.000
		df	0	6274
	attention	correlation	-.052	1.000
		significance	.000	.
		df	6274	0

Number of follows and attention are factors which both belong to sequencing variables to measure the influence and vitality of users from different perspective, so we chose the Spearman correlation coefficient to analysis the degree of close between each other. Analysis results are as shown in the above table, in which we can see the correlation coefficient of number of follows and attention is 0.052 after removing the influence of number of contents. According to the definition of correlation coefficient on the above, we can conclude that the two variables are weakly related, which also can be said that the number of follows and attention of a user are not in a positive correlation relationship.

Table 5. Correlation coefficient of number of follows and contents using number of attention as control variables

control variables			follows	contents
attention	follows	correlation	1.000	.186
		significance	.	.000
		df	0	6274
	contents	correlation	.186	1.000
		significance	.000	.
		df	6274	0

It can be seen from the above table that the correlation coefficient of number of follows and contents is 0.186 after removing the influence of number of attention. According to the definition of correlation coefficient on the above, we can conclude that the two variables are weakly related, which also can be said that the number of a user's follows

and contents are not in positive correlation relationship.

Table 6. Correlation coefficient of three variables about ordinary uses

			follows	attention	contents
Spearman rho	follows	correlation	1.000	.309**	.631**
		Sig.	.	.000	.000
		N	6277	6277	6277
	attention	correlation	.309**	1.000	.505**
		Sig.	.000	.	.000
		N	6277	6277	6277
	contents	correlation	.631**	.505**	1.000
		Sig.	.000	.000	.
		N	6277	6277	6277

From the table, we can see that the correlation coefficient of number of follows and attention is 0.309. According to the definition of correlation coefficient in the above, we can conclude that the two variables are lowly related, but there is a certain correlation. Combining with the above analysis, we can draw a conclusion that there exists relationship between number of follows and number of attention only when under the influence of number of content. That is to say, if a user often releases twitter, he will cause more attention, which will increase his number of follows, and in turn the number of his attention improves. The improvement of number of follows of a user will drive him pay more attention to his follows, and thus the number of his attention will increase. And if a user has very few published twitter, he will not cause the interest of others even if he pay a lot attention.

The correlation coefficient of number of follows and number of contents of users is 0.631. According to the definition of correlation coefficient on the above, we can conclude that the two variables are moderately related. Combining with the above analysis, we can draw a conclusion that there exists relationship between number of follows and attention only when they under the influence of number of attention. That is to say, for an ordinary user, if he pays little attention to others, the influence of the number of follows is extremely weak no matter how many twitters he releases.

Table 7. Correlation coefficient of three variables about authenticated uses

			follows	attention	contents
Spearman rho	follows	correlation	1.000	-.033**	.290**
		Sig.	.	.000	.000
		N	12250	12250	12250
	attention	correlation	-.033**	1.000	.298**
		Sig.	.000	.	.000
		N	12250	12250	12250
	contents	correlation	.290**	.298**	1.000
		Sig.	.000	.000	.
		N	12250	12250	12250

The author carries on correlation analysis of the number of follows, attention and contents using SPSS software. The result is shown in the above table, from which we can see that the correlation coefficients are respectively -0.033, 0.290 and 0.298. According to the definition of correlation coefficient on the above, we can conclude that this three variables are weakly related, which can also be said that there is no obvious linear relationship between the three variables. It means that the various values are disturbed to a large extent under the influence of celebrity charm.

3.3. Clustering analysis

3.3.1. Ordinary users clustering analysis: In this paper, 6277 normal users are divided into four groups randomly. We use SPSS software to do clustering analysis on each group respectively. At last, we put all the users together and make analysis on the thesis of K-means clustering algorithm. Number of follows, attention, contents and mutual are selected as variables involved in clustering. The result shows that the users are divided into four types eventually, clustering result is as shown in the figure below.

Table 8. Clustering analysis and result of ordinary users in group 1

	clustering			
	1	2	3	4
follows	6669	419162	946945	6115853
attention	701	407	335	1039
contents	2357	4526	2262	3876
mutual fans	297	295	193	1036

cluster	1	1527.000
	2	32.000
	3	6.000
	4	4.000

Table 9. Clustering Analysis and Result of ordinary users in group 2

	clustering			
	1	2	3	4
follows	6827	5824941	3939140	13657754
attention	579	254	511	80
contents	1479	1128	51256	1451
mutual fans	121	201	125	73

cluster	1	1557.000
	2	5.000
	3	2.000
	4	6.000

Table 10. Clustering analysis and result of ordinary users in group 3

	clustering			
	1	2	3	4
follows	28580	952343	2198869	4748299
attention	432	413	469	948
contents	1729	6694	5695	4367
mutual fans	113	150	165	354

cluster	1	1460.000
	2	88.000
	3	19.000
	4	3.000

Table 11. Clustering analysis and result of ordinary users in group 4

	clustering			
	1	2	3	4
follows	21283	1523525	5341515	10594966
attention	395	592	1285	958
contents	1640	5846	7895	52651
mutual fans	177	225	1048	339

cluster	1	1526.000
	2	37.000
	3	2.000
	4	3.000

Table 12. The ordinary user clustering analysis and cluster grouping result

	clustering			
	1	2	3	4
follows	43572	4600027	10594966	13657754
attention	526	626	958	80
contents	1914	8515	52651	1451
mutual fans	178	385	339	73

cluster	1	6244.000
	2	24.000
	3	3.000
	4	6.000

We can conclude that, no matter how many members are there in the tests, ordinary users can be divided into the following several types in each group and total table according to the results of system clustering:

1) Social type

Such users are the long tail part in the Long Tail Theory. Although they have low value in all kinds of figures, they are the most common users in Micro-blog. By the result of clustering we can see that the general features of the users are few fans, followers, contents and higher mutual relationship. The main activity of such users on Micro-blog is social activity. On one hand, they communicate with friends and other users of

Micro-blog; on the other hand, they get interesting information from these massive data. They tend to make friends with those who share the same interests. According to the large amount of users of this type, they are becoming the backbones in the transmission of topics. Usually, people can tell the hot of topics from the diffusion degree among this type of users.

2) Informative type

The characteristic of this kind of users is that their performance is between ordinary social users and other types of users in all kinds of properties except the most mutual fans. They possess more follows, attention, contents and mutual fans compared with users of social type. Besides, they tend to make friends who share the same interest in addition to real life friends. They usually have their own small communities in which contents mainly focus on a particular direction. This kind of communities are often full of vitality, which makes the information spread easily. Normally, this kind of users assumes the same role with users of social type in the process of topic transmission, which also means the final node in a community of topic. There are also some users who can be regarded as the center of a certain community. They are the key to maintain contact between members of the community. This kind of users can take important roles in the topic propagation sometimes, which can be said that they are the center node of a community of topic.

3) Active type

This kind of users usually has the highest value in various characteristics except number of mutual fans compared with users of other types. The cluster result shows that there are little amount of this kind users, but they are likely to be involved in all kinds of interaction on Micro-blog platform. Their major activities are obtaining and sharing information, which often make them the second source of all sorts of topics. Such users, of course, also have certain social demand, but the demand is not very important compared with the activeness of releasing information. Usually, one can get access to the latest news on Micro-blog platform timely if pay attention to this kind of users. This type of users generally located in the middle nodes in topic propagation.

4) Grassroots celebrity type

Their general characteristics are the most number of followers, fewer attention, mutual fans and contents. One of the most obvious differences between active users and ordinary celebrities is that the contents and followers of grassroots celebrity are much lower than active users. The most important characteristic is that the influence of this kind of users is not below authenticated users. Their activities on Micro-blog platform are not very active, but they usually get a lot of attention. They often become an important node in the process of the propagation of a hot topic because of the high value they create. This type of users generally located in the center within certain limits in topic propagation.

3.3.2 Certificate users clustering analysis: Using SPSS to do K-Means average clustering of 12250 certificate users, we found that clustering of 12250 certificated users is not obvious and data between class and class difference is small. Therefore, we believe that the clustering is invalid. In the transmission process of hot topic, we will treat authenticated users as a category, and give them the same weight on the influence in the process of the propagation of topic.

3.4. Discriminate analysis

Through the K-Means average clustering, Micro-blog ordinary users can be divided into social, informative, active and grassroots celebrity types in this paper. In the process of topic propagation analysis, we need to decide which kind of type the users are. In accordance with the above data, Discrimination function was established based on Fisher discriminate method. Coefficients are shown in the table below:

Table 13. Classification function coefficient table

	category			
	1	2	3	4
follows	6.333E-7	7.265E-5	.000	.000
attention	.002	.003	.005	.008
constants	2.189E-5	.000	.001	-.002
mutual fans	.000	-.002	-.014	-.011
(constant)	-1.892	-152.351	-961.772	-1640.184

According to the established discriminate function, we have carried on the part of the original data, results show that the discriminate function effect is good.

4. The empirical analysis

In order to verify the rationality of the classification and the accuracy of the discriminate function, the author use the PKUVIS software to track the spread of the blog about the lost of Malaysia airlines flight MH370 released on March 8th visually. Micro-blog visual analysis tools can show the propagation process of an event in clearly intuitive view. With the help of Micro-blog visualization tools, we can quickly find key figure, key blog and important point of view and analyze the occurrence and development process of the event. The image below shows the topic of transmission in circle view.

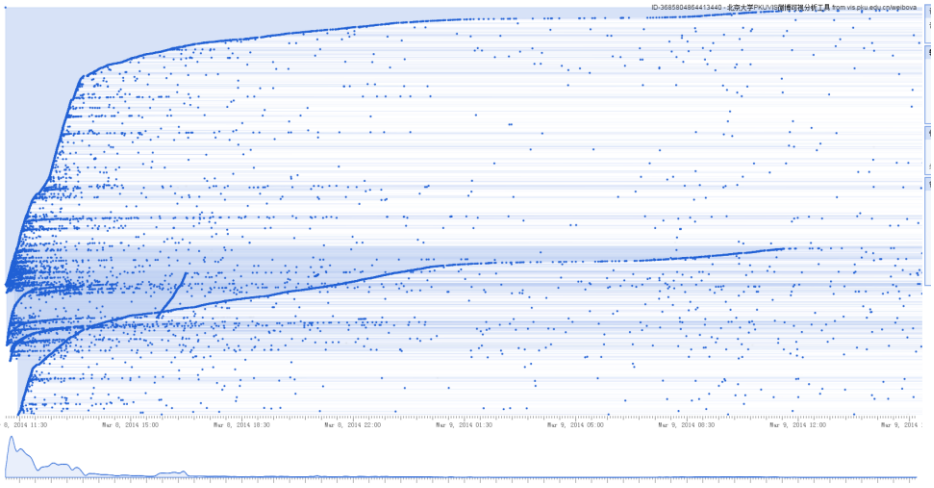


Figure 1. Visualization of circle view in topic propagation

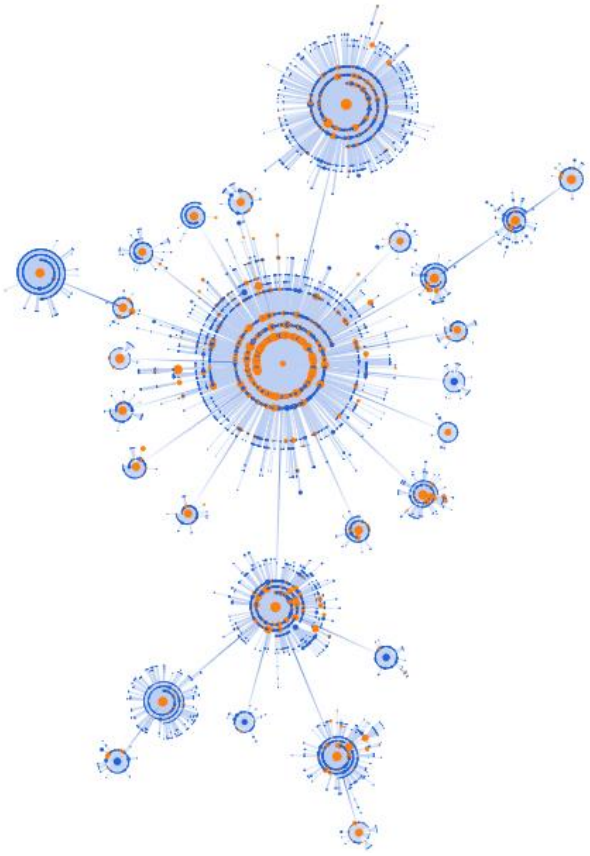


Figure 2. Visualization of sail shape view in topic propagation

Every node in Figure 1 represent a blog, attachment represent relations between the users. In the above figure the core of the largest circle is Malaysia airlines, which is also the source of the topic. We can find that the spread of topics in Micro-blog are generally radiating by the graph, spreading outward via a central point, but there will be some multiple small circles in the transmission of topic, which can also be called subject community. The spread of the topic transmission process can be summarized as community diffusion. In order to see the authenticated user role in the process of the propagation of topic more clearly, the author carried on the different colors to authenticate the user node's tag. The result is shown in figure 2. In this topic, ordinary users are to 18230, accounting for 95% of the total number of; authenticated users are to 890, accounting for 5% of the total number. This proved the long tail theory in Micro-blog Once again. It can be seen in the figure that the authenticated users are in the center of much propagation process in a circle of topic. They played a great promoting role in the spread of topics. Besides, some ordinary users are also in the center of the small community, which has become the second source of topic.

Analysis the transmission process of the topic, we can find that besides topic source Malaysia airlines, most users which are forwarded directly the most are authenticated. The author took part of the ordinary users located in the center of the topic of community information, which are shown in the table below:

Table 14. Part of the user information in topic transmission

	follows	attention	contents	mutual
User 1	640321	200	4780	42
User 2	836288	223	1451	38

Put the information of users into the discriminate separately, we can draw that users 1 and 2 are informative users. The results are in line with the clustering analysis of the characteristics of the loyal users. According to the observation of the node user information directly, we find that the vast majority of nodes users belong to the social users. The analysis of the above on the clustering result is verified.

5. Conclusion

Using Sina Micro-blog platform, through analysis of a large number of users data, the paper mainly get the following conclusions.

There exists huge difference between ordinary users and authenticated users on their characteristics. For ordinary users, the correlation analysis shows that there is a certain linear relationship between all their characteristics. Ordinary users can take action to change the corresponding attribute values. For authenticated users, the correlation analysis shows that their attribute values are more influenced by other factors, so there is no obvious linear relationship. According to the clustering results, ordinary users are divided into four different types of social, informative, active and grassroots celebrity for the different behavior characteristics of the user clustering analysis. Type of social users and grassroots celebrity users are most likely to center nodes in a topic community. Active users, who are important intermediate node within the topic community, generally promote the spread of topics. Social users, who are the end of topic propagation, most located in the final nodes in the topic communities. The result of clustering of authenticated users is not obvious, therefore, authenticated users , assuming the same role in topic propagation process, can be considered to the same class. Using visualization tools to track the spread of twitter about the lost of Malaysia airlines flight MH370, the author simulate the diffusion of the topic and analyzes the results. Results show that the above classification of users is reasonable

Research of this paper is based on a large number of user data, but compared with the total numbers of users in Sina Micro-blog, it's still insufficient. Besides the data are only from Sina Micro-blog platform, which did not consider different platform users behavioral characteristics fully.

Acknowledgements

This research was funded by Beijing Science and Technology Committee.

Project Number is Z131100005613017, Named by The Construction and Demonstration Applications of Medical Association Model of Critical Illness Collaborative Control in Health Care Reform of BeiJing.

References

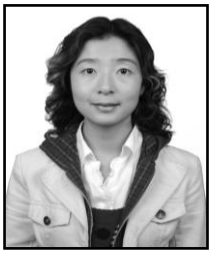
[1] Da Yuan, Richen Liu, Xiaoru Yuan, Seismic Visualization, [J]. Journal of Computer-Aided Design & Computer Graphics, 27(1):36-46 **(2015)**.
[2] Zuchao Wang, Xiaoru Yuan. Visual Analysis of Trajectory Data, [J]. Journal of Computer-Aided Design & Computer Graphics, 27(1):9-25, **(2015)**.
[3] DAI Guozhong, CHEN Wei, HONG Wenxue, LIU Shixia, QU Huamin, YUAN Xiaoru, ZHANG Jiawan

and ZHANG Kang. Information Visualization and Visual Analytics: Challenges and Opportunities, [J]. Scientia Sinica Informationis, 43(1): 178-184, **(2013)**.
[4] S.Yardi, D.Boyd. Dynamic Debates.An Analysis of Group Polarization Over Time on Twitter[J]. Bulletin of Science, Technology&Society, 10, **(2010).**
[5] Xiao Qiang, Zhu Qinghua. Study Of Micro-blog User Character And Types[J]. Information Science.31[12]:69-74, **(2013)**.
[6] Wang Xiaoguang. Empirical Analysis On Behavior Characteristics and Relation Characteristics Of Micro-blog Users[J].Library And Information Service,54(14):66-70, **(2013)**.
[7] J. Tang, Z. Liu, M. Sun, and J. Liu. Portraying user life status from micro blogging posts[J]. Tsinghua Science and Technology, 18(2):182–195, **(2013)**.
[8] A. M. MacEachren, A. Jaiswal, A. C. Robinson, S. Pezanowski,A. Savelyev, P. Mitra, X. Zhang, and J. Blanford. SensePlace2:Geo Twitter analytic support for institutional awareness. In Proceedings of IEEE Conference on Visual Analytic Science and Technology,181 –190, **(2011)**.
[9] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In Proceedings of the 19th international conference on World Wide Web, 851–860, **(2010)**.
[10] J. Ratkiewicz, M. Conover, M. Meiss, B. Gonc¸alves, S. Patil, A. Flammini, and F. Menczer. Truthy: mapping the spread of AstroTurf in micro blog streams. In Proceedings of the 20th international conference companion on World Wide Web, 249–252, **(2011)**.
[11] Kimura M, Saito K, Nakano R. Extracting influential nodes for information diffusion on social network[C]. Proceedings of the 22nd AAAI Conference on Artificial Intelligence. AAI,1371—1376, **(2014)**.
[12] L. Yu, S. Asur, and B. A. Huberman. Artificial inflation: The true story of trends in sina weibo. CoRR, **(2012)**.
[13] Y. Qu, C. Huang, P. Zhang, and J. Zhang. Microblogging after a major disaster in china: a case study of the 2010 yushu earthquake. In Proceedings of the ACM conference on Computer Supported Cooperative Work, 25–34，**(2011)**.
[14] F. B. Viegas, M. Wattenberg, and K. Dave. Studying cooperation and conflict between authors with history flow visualizations. In Proceedings of the SIGCHI conference on Human Factors in Computing Systems,575–582, **(2014)**.
[15] U. Brandes, P. Kenis, J. Lerner, and D. van Raaij. Network analysis of collaboration structure in Wikipedia. In Proceedings of the 18th international conference on World Wide Web, 731–740, **(2009)**.
[16] D. Zhai, J. Yu, F. Gao, L. Yu and F. Ding, K-means text clustering algorithm based on initial cluster centers selection according to maximum distance, [J],Application Research of Computer, 31(3): 713-715, **(2014)**.
[17] C. S. Li, X. L. An and R. H. Li, "A chaos embedded GSA-SVM hybrid system for classification", Neural Computing and Applications, 26(3): 713-721, **(2015)**.
[18] T. E. Turker, T. Cumhur and C. Merve, "A wrapper-based approach for feature selection and classification of major depressive disorder-bipolar disorders", Computers in Biology and Medicine, 46: 127-137, **(2015)**.
[19] A. J. G. Mohammad, K. Shamsul, M. Faruq and J. F. Hessam, "Feature decision-making ant colony optimization system for an automated recognition of plant species", Expert Systems with Applications, 42(5): 2361-2370. **(2015)**.
[20] S. Palanisamy and S. Kanmani, "Classifier ensemble design using artificial bee colony based feature selection", International Journal of Computer Science, 9(3):522-529, **(2012)**.

Authors

QiuLi Qin, she is an associate professor at Beijing Jiaotong University. She received her Ph.D degree in 2002. Her main research interests include enterprise informatization, cloud computing and data mining.

Xing Yang, she is currently persuing her Master degree from Beijing Jiaotong University. Her main research interests include cloud computing and data mining.



Hua Gu, she is currently persuing her Ph.D degree from Beijing Jiaotong University. Her main research interests include cloud computing and data mining.

