International Journal of Automation and Computing 11(2), April 2014, 119-127
DOI: 10.1007/s11633-014-0774-9

Big Data Modeling and Analysis of Microblog Ecosystem

Hua-Ping Zhang[1] Rui-Qi Zhang[1] Yan-Ping Zhao[2] Bao-Jun Ma[3]

[1]School of Computer Science, Beijing Institute of Technology, Beijing 100081, China
[2]School of Management and Economics, Beijing Institute of Technology, Beijing 100081, China
[3]School of Economics and Management, Beijing University of Posts and Telecommunications, Beijing 100876, China

Abstract: Recent progress of Web 2.0 applications has witnessed the rapid development of microblog in China, which has already been one of the most important ways for online communications, especially on sharing information. This paper tries to make an in-depth investigation on the big data modeling and analysis of microblog ecosystem in China by using a real dataset containing over 17 million records of SinaWeibo users. First, we present the detailed geography, gender, authentication, education and age analysis of microblog users in this dataset. Then we conduct the numerical features distribution analysis, propose the user influence formula and calculate the influences for different kinds of microblog users. Finally, user content intention analysis is performed to reveal users′ most concerns in their daily life.

Keywords: Microblog mining, user activity pattern, user personalization, user intention, microblog ecosystem.

1 Introduction

Recent progress of Web 2.0 applications has witnessed the rapid development of microblog in China (i.e., SinaWeibo), which has already been one of the most important ways for people′s online communications, especially on sharing information[1]. Since its launch in August 2009, SinaWeibo has grown into the biggest Chinese microblog with 500 million registered users by the end of 2012[2]. In SinaWeibo, there are more than 46.2 million active users and over 100 millions posts issued each day[3]. Similar to other famous microbog platforms, such as Twitter, any registered user on SinaWeibo can conduct two kinds of activities. The first kind lies in self-activities, such as expressing his own status, location or emotion in a post within a limit of 140 characters. The other kind of activities are associated with other users, such as following others, commenting or retweeting on existing posts. There are several special character symbols on SinaWeibo to represent certain meanings. For instance, "#" usually represents a topic, while "@" may indicate making comments or retweeting.

With users′ online activities being more and more important in their daily life, it is very important and valuable for both operators and researchers to understand users′ behavior in Weibo[4]. First, system operators can make better site design when they know who the users are. Second, knowing why the users use Weibo will make the system operators provide suitable services. Third, if we know the user behavior well in Weibo, we may infer what will happen in the future in the real life. Moreover, Weibo grows faster, and users of Weibo enjoy a different culture, because it is only used by Chinese. Tweets in Weibo can contain images and videos besides text message and links, and dealing with every tweet′s reply and comment is different in Weibo and Twitter. All these reasons make us want to know the distinct characteristics of Weibo.

Would you want to make a better understanding about Chinese people? How are the distributions of Weibo users on geography, gender, authentication levels, education or age? How are the different influences between various kinds of users? And what is the most important issue in most people′s hearts and minds? To answer the above questions, it is necessary and important to make an in-depth investigation on SinaWeibo.

In this research, we first propose the definition of "microblog ecosystem", which is regarded as an organism that incorporates microbolg users, their posts and all of their online activities together. It has some basic statistical features, numerical features as well as context features. Our work in this article aims to conduct an in-depth analysis of the specific Chinese microblog ecosystem (i.e., SinaWeibo) in the macro perspective of big data by using a dataset containing more than 17 million records of Weibo users. Although the register number of SinaWeibo has reached 500 millions, among those there are just 40% users who choose to fill their profile information and there are many inactive users, machine generated users and zombie users. We filtered out all those useless users, then made the collection of more than 15 gigabyte data from 17 million active real records.

The remainder of this paper is organized as follows. Section 2 introduces related research work. Section 3 describes the details about our data crawled on SinaWeibo. And the basic feature analysis of microblog ecosystem is presented in Section 4. Based on this analysis, we conduct numeric feature distribution for user interactive behaviors and propose a novel method to reflect users′ influence in Section 5. In Section 6, user intent analysis based on text content will be discussed in detail. Finally, conclusions are outlined in Section 7.

2 Related work

In the related research area, the following studies have made big contributions and provided important guides. Guo et al. made a comparative study of users′ behavior

Manuscript received July 31, 2013; revised November 21, 2013
This work was partly supported by National Natural Science Foundation of China (No. 61272362), National Basic Research Program of China (973 Program) (No. 2013CB329606), and High-Tech Development Plan of Xinjiang (No. 201212124).

between SinaWeibo and Twitter[4]. Java et al.[5], Mislove et al.[6], and Corbett[7] discussed demographic analyses of mainly social network websites like Twitter and Facebook, including their user distributions of geographic, gender, race and so on, with 54 million world records between 2006–2009. They also gave the growth trends and network properties of Twitter and proposed a method to find communities. Unlike the studies mentioned above, our work focuses on Chinese microblog and SinaWeibo, over 17 million records, more than 15 gigabyte data, between 2009–2013, to study the corresponding properties of Chinese social network. We present our analyses of user distributions which were rarely studied before. Inspired by the research of Kwak et al.[8], Yan et al.[9], Barabási[10] and Tsur and Rappoport[11] discussed followers and following distributions, which are also known as in-degree and out-degree distributions of users, fitting the famous power-law distribution of social network. We make similar research on SinaWeibo using the linear regression method which was also applied by Tsur and Rappoport[11], but found the distribution does not fit power-law pretty well. Mangai et al.[12] and Balamurugan and Rajaram[13] gave some useful feature selection methods to process large scale data, which provided much valuable references for our data processing.

Another interesting research is users′ influence. Bakshy et al.[14] discussed users′ influence by constructing spread tree while Cha et al.[15] measured users′ influence by using numbers of users′ followers and transfers and Quercia et al.[16] calculated users′ influence with considering users′ posts, comments, transfers, mentions and followers. As Romero et al.[17] proposed an passive-influence algorithm with consideration of both influence and passivity, we gave a formula to calculate users′ influence in SinaWeibo by considering users′ posts, followers, followings, and follow-back rate by referring to the research of Yang et al.[18] who discovered that only 25.5% of all the information was generated by transfer via studying the transfer mechanism in Twitter. Users′ behavior and users′ intention were discussed in a variety of aspects, such as Lee et al.[19] tried to mine users′ behavior pattern by marking users′ geography information via mobile microblog, while Lingad et al.[20] used named entity recognizers to extract local information disaster-related microblogs and Kwak et al.[21] discussed the dynamics of the behavior known as “unfollow” in Twitter and discovered the major factors affecting the decision to unfollow. We analyzed users′ intention by analyzing their microblog text and got interesting results of real big data of SinaWeibo with many first-hand new discoveries other than the small dataset results with traditional techniques, but natural language process (NLP) and big data analysis on the tags or semantic contents and contexts.

3 Dataset description

In this research, we mainly focus on the users′ profile information, which is normally published to public. By using the network spider, we totally crawled over 17 million records of users′ profiles randomly from the beginning of the platform service to the May 1st, 2013 after screening out the inactive users. In each record, there are usually more than ten fields or attributes to represent the detailed information of each Weibo user, such as user ID, URL, name, gender, birthday, address, fansNum, summary, wbNum, gzNum, Blog, realname, email, QQ and so on. Table 1 lists and describes the attributes we analyze in this research. Although our dataset may not cover the full users′ profile information on SinaWeibo, we posit that it can still reveal certain trends and some interesting features of users from the perspective of big data due to the randomness of crawling.

Table 1 Attributes of user profile analyzed in this research

Attribute name	Description
ID	User ID
gender	Gender information
wbNum	Number of tweets
fansNum	Number of followers
gzNum	Number of followings
summary	Self-introduction of users
address	Location information
edu	Education information
renZh	Authentication level

Normally, for each user, the ID and gender information are mandatory. Meanwhile, since users may not provide the corresponding information, one or more attributes would not be crawled and display the null value in some cases. For instance, in the total 17 million records, the filled (i.e., not null) attribute of education information only accounts for around 3% of all the users (see Fig. 1).

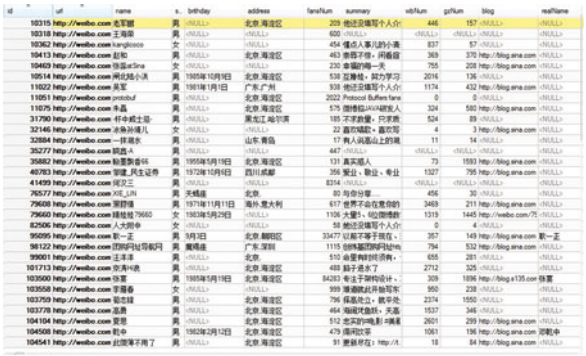

Fig. 1 Snapshot of dataset

4 Basic statistic features analysis

In this section, we will begin our analysis based on the dataset we collected from SinaWeibo with some basic features statistics, such as geography, gender, authentication, education and age analysis. From the analysis, we could find the answers to the following questions in the subsequent subsections:

1) Which districts own the highest and lowest user density of SinaWeibo?

2) What is the relationship between the average gross domestic product (GDP) and the microbolg users?

3) How is the distribution difference between male and female Weibo users?

4) How is the distribution situation of users with different authentication levels?

5) How about the education and age distribution of Weibo users?

4.1 Geography analysis

For more than 17 million records in our dataset, around 16.5 million users have specified locations information. Fig. 2 displays the distribution of user density for each province in China. The vertical axis represents the number of Weibo users in one thousand people, in which the provincial population data could be found in the statistical yearbook of China government. From the histogram in Fig. 2, it could be found that the average user density is 10.8 users per thousand people. Moreover, Beijing has the highest user density with 79 users per thousand people while Gansu province possesses the lowest user density (i.e., 3.9 users per thousand people). Furthermore, there are 12 out of 34 provinces with user density above the average level.

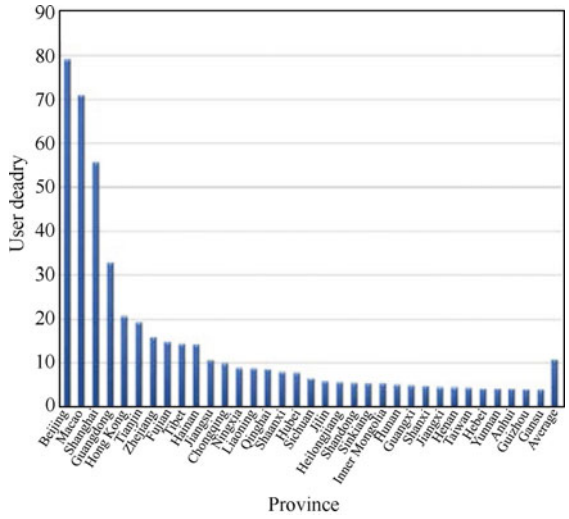

Fig. 2 User density distribution of each province

One may argue from Fig. 2 that the user density in each province has a close relationship with its population. However, as we analyzed, the above distribution of user density may fit the developing levels of Chinese districts, implying that the more developed one district, the more Weibo users. Other important influential factors may lie in population, economic scale and so on. It is posited that the density of Chinese microblog users can be considered as the popularizing rate of microblog. Only with computers, Internet, smart phones and other smart devices becoming popular, it is possible that more people will use microblog, which will firstly occur in relative modern areas.

In addition, it is easy to make a comparison between the user density and the average GDP per person for each province. Fig. 3 shows the distribution of average GDP per person in each province in China. It is found that the distribution in Fig. 3 has certain similarities to that of user density in Fig. 2. We notice that to some extent, the higher province's average GDP, the more Weibo users, which demonstrates that microblog users density of one district would reflect the developing level of its economy to some degree.

Furthermore, we also conduct the linear regression between average GDP and user density, in which the user density is regarded as the dependent variable (y) and the average GDP is treated as the independent variable (x). The regression result is illustrated as

$$y = 0.86413x + 5.396. \tag{1}$$

Meantime, it is found that the coefficient of average GDP (i.e., 0.86413) is statistically significant and the R square value of the regression model is 38.75%, which implies that the average GDP could account for nearly 40% of the impacts on user density.

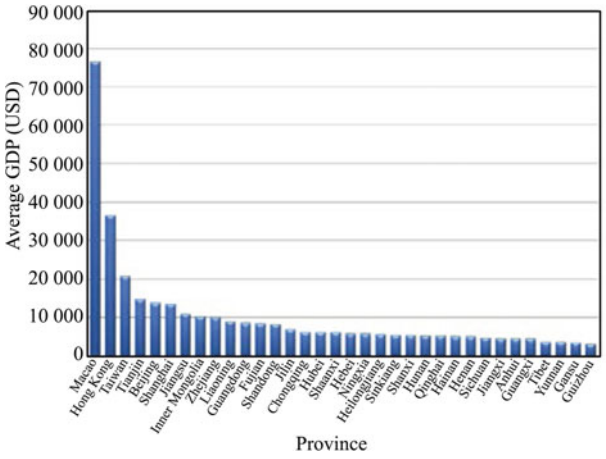

Fig. 3 Distribution of average GDP

4.2 Gender analysis

Gender analysis shows the distribution of male and female users. Fig. 4 displays that male users occupy 45 percentages while female users reach 55 percentages.

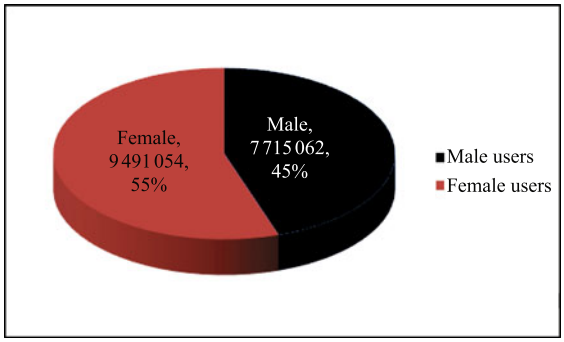

Fig. 4 Distribution of male and female users

It is easy to find that the number of female users is obviously larger than that of male users. Why does it occur? Why is the proportion of female and male not close to 1:1? As we analyzed, one possible explanation to this result may be that the jobs of females and males are rather different that females may have more spared time to spend on microblog.

4.3 Authentication analysis

The "renZh" field represents whether each user has gained an authentication or not as well as the corresponding authentication level. There are totally five authentication levels in our datasets described in Table 2.

Table 2 Levels of authentication

Authentication level	Number	Description
No authentication	16 746 493	97.26% in total
Level 1	201 423	Authenticated individual (1.17%)
Level 2	23 150	Government media, etc. (0.13%)
Level 3	266 428	Enterprises, etc. (1.43%)
Level 4	931	Experts and elites (0.01%)

Authenticated users are normally VIP users who own relative high authority. Though the proportion of these users is only less 3% in total, they actually pose the most prominent influences in the microblog social network. It is found that authenticated users usually have some common characteristics. For instance, they all own large numbers of followers and they always become the information centers or sources of many hot discussion topics on SinaWeibo. With the small-world characteristic[7], these authentication users in microblog social network will usually be the key nodes of the path of users′ map.

4.4 Education and age analysis

In our dataset, over 662 000 users registered their education information, which only accounts for 3.8% of all the users. Among them, 83.2%, i.e., around 551 000 users, are graduates or undergraduates.

Meanwhile, Fig. 5 shows the age distribution of users. It clearly demonstrates that most of the Weibo users are young people. Users with age ranging from 21 to 40 have reached a percentage of over 75%. It is easy to explain that young adults are more willing to accept new things than older people.

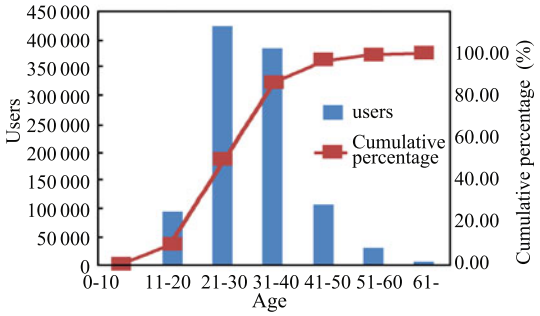

Fig. 5 User age distribution

5 Numerical-features distribution and influence modeling

Each user has some numerical features in his profile information, such as posts number, follower number as well as following number. In this research, we treat these numerical features as variables of one user, which will reflect one′s active degree and his influential power in the microblog network. To analyze one′s numerical features will obtain the quantitative characteristics of each user.

In addition, numerical characteristics analysis is also one important part of our research model, which will show the panoramic view in a numerical perspective and disclose the distributions of users′ posts, followers and followings, respectively. After that, the prediction could be conducted based on the regression models.

In detail, our model of microblog ecosystem will be then demonstrated by the analysis of posts-users, followers-users, followings-users and context. Apart from text analysis which will be shown in the next section, we focus on discussions of numerical features and user influence analysis in this section.

It is noticed that in order to analyze our data and show results more conveniently, we draw the points in the double logarithmic coordinates.

5.1 Posts-users analysis

The number of published posts could show the activity degree of users from one perspective. To some extent, the more posts published, the more active a user is. For instance, in our dataset, the user with ID of 10 057 has published the largest number of posts, reaching 613 070 posts. Fig. 6 shows the distribution of post number and user number.

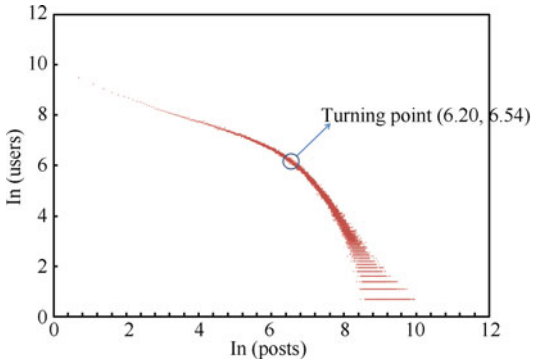

Fig. 6 Distribution of post number and user number

From Fig. 6, it is obviously found that the number of posts and its correspondent user number do not fit a line, which means that the distributions of them do not follow the power law. The inflation or turning point is (6.2, 6.54), which divides the discrete points into two parts. Specifically, the points in each part can fit a line well, as shown in Figs. 7 and 8, respectively. Therefore, we call this kind of distributions of posts and users following the "piecewise power law".

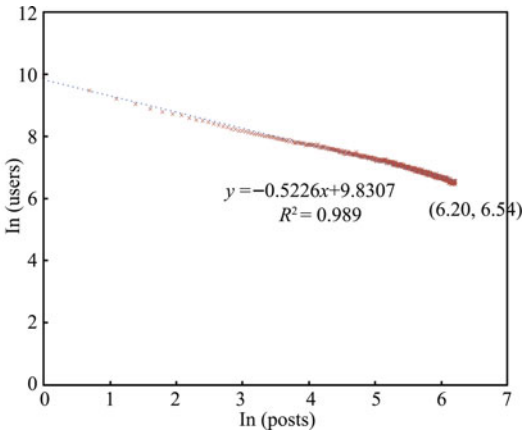

Fig. 7 First part linear regression of posts-users

In Fig. 7, the discrete points regress to the line

$$y = -0.5226x + 9.8307 \tag{2}$$

which means that this part fits the power law

$$y = 18\,595.97x^{-0.5226}. \tag{3}$$

In the meantime, the discrete points in Fig. 8 regress to the line

$$y = -1.9771x + 19.04 \tag{4}$$

demonstrating that this part fits the power law

$$y = 185\,766\,301.8x^{-1.9771}. \tag{5}$$

It is found that these two parts have different power values.

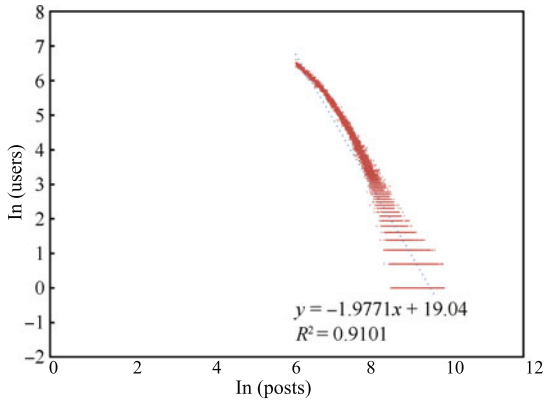

Fig. 8 Second part linear regression of posts-users

5.2 Followers-users analysis

In this subsection, the followers-users distribution analysis will be given, which is shown in Fig. 9 and reflects the analogous distribution to that of posts-users.

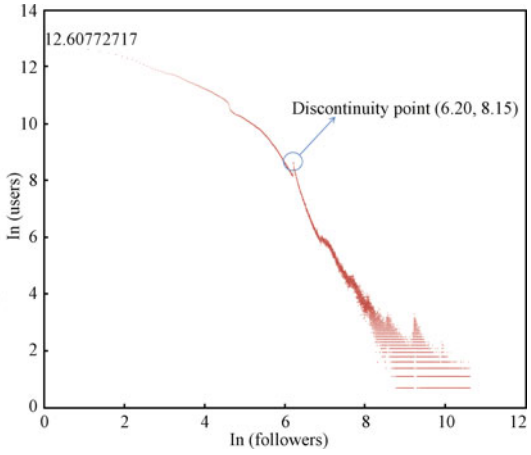

Fig. 9 Distribution of followers and users

Similar to Fig. 6, the curve in Fig. 9 also fits the "piecewise power law". Specifically, this curve is divided into two parts by a discontinuity point (6.2, 8.15). The discrete points of the first part regress to the line as displayed in Fig. 10.

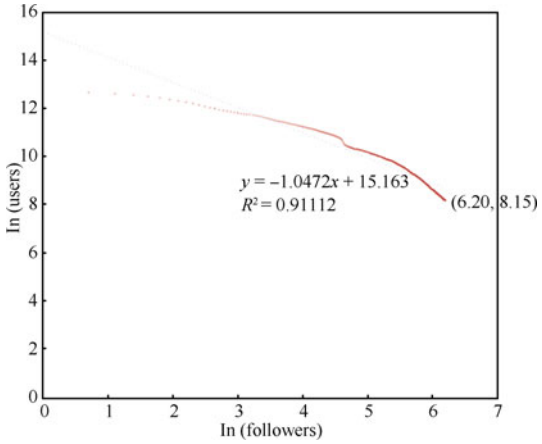

Fig. 10 First part linear regression of followers-users

$$y = -0.592x + 13.286 \tag{6}$$

which is corresponding to

$$y = 588\,893.1x^{-0.5192}. \tag{7}$$

The second part regresses to the line, as displayed in Fig. 11:

$$y = -1.5214x + 15.775 \tag{8}$$

and its power law equation is

$$y = 7\,095\,703x^{-1.5214}. \tag{9}$$

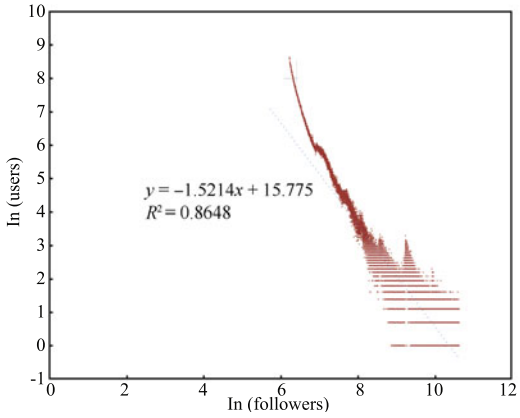

Fig. 11 Second part linear regression of followers-users

It is obviously that the curve in Fig. 9 is not as smooth as that of posts-users distribution in Fig. 6, in which the former has several break points. The most important one is point (6.2, 8.15) that is already mentioned above. Considering Figs. 6 and 9 together, it is found that the horizontal coordinates of the break-points in both figures have the same value, namely 6.2, which means that users whose tweets number is less than 496 or whose followers number is below 496 will be classified into the first power law situation and the rest fit the second piecewise power law.

By comparing the equations of the first power laws of posts-users and followers-users (i.e., (3) and (7)), we find that their power values are very close to each other. In detail, the power value of posts-users is −0.5226 and that of followers-users is −0.5192, i.e., both are rather close to the value of −0.520.

5.3 Followings-users analysis

Similarly, we obtain the distribution of followings-users, which is shown is Fig. 12.

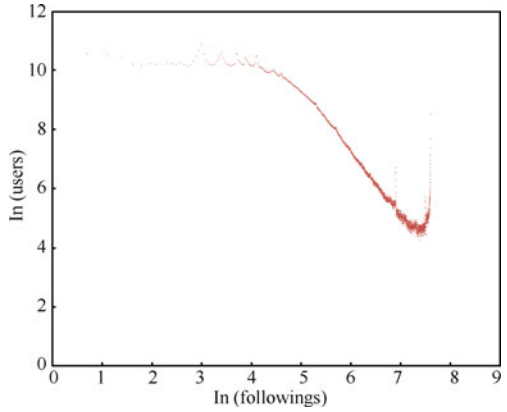

Fig. 12 Distribution of followings-users

This distribution is much more different from those of the followers-users and posts-users. By ignoring the discrete points at the beginning with a smaller followings number, most the other points fit the power law well, as displayed in Fig. 13. The equation of the line is

$$y = -1.9541x + 18.94. \tag{10}$$

Thus the power law equation is

$$y = 168\,088\,301x^{-1.9541}. \tag{11}$$

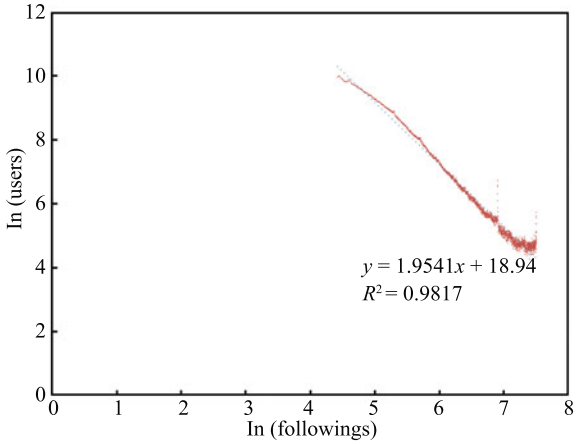

Fig. 13 Distribution of followings-users without discrete points at the beginning

Because of the limitation of the max following number that each user can only follow no more than 2 000 users, many points disperse near the vertical line $x = 7.6$, where users′ followings reach the max number.

5.4 User influence analysis

Users′ influence is a very important topic in microblog research[22, 23]. In this subsection, we propose our influence computation formula for each user

$$\text{Influence}(\alpha) = \frac{(\#\text{followers} - \alpha \cdot \#\text{following})}{\#\text{posts}} \tag{12}$$

where #followers means the user′s followers number, #following is the user′s followings number and #posts represents the user′s posts number. Moreover, α is the follow-back rate, which represents the probability that a user′s following follows him.

It is assumed that the more followers a user has, the more influential power he owns. It is obvious that if a person follows a user, it has some probability this user follow-back that person. Thus we multiply a follow-back rate by #following. Dividing by #posts, we can obtain the user′s influence per post. In this research, we calculated the average influence values for different kinds of users with three distinct α values (i.e., 1, 0.5 and 0.3), such as V users, NV users, male users, female users, V male users, V female users and so on. Table 3 lists the characteristics and influences for different kinds of users.

First, from Table 3, it is revealed that on average V users have much more followers, followings and issued more posts than NV users. In addition, (V) male users own less posts and more followers as well as followings than (V) female users. Furthermore, it is found that V users possess outstanding advantages over NV users on influential power and (V) male users are nearly twice influential than (V) female users. With regarding to the impact of the follow-back rate α, the influences of all kinds of users increase with the decrease of the follow-back rate, which is consistent with the influence computation equality (12).

6 User intent analysis

User intent analysis is another important perspective to provide a glance to the whole ecosystem of microblog[1, 5]. Since users′ self-introductions are stored in the summary field, we could analyze this field to complete our content analysis. The summary field may reveal users′ emotions, attentions, hobbies and so on. Through mining the content of users′ self-introductions, we could understand and learn more about the users′ concerns, thoughts and needs in depth.

Table 3 Influence of users in our research

	All users	V users	NV users	Male	Female	V male	V female
Average followers	512.23	6965.11	337.23	600.75	440.42	7887.36	5833.1
Average posts	774.92	1435.41	704.89	685.26	854.06	1202.99	1524.11
Average followees	176.66	342.19	171.08	181.32	172.41	362.67	313.81
Influence ($\alpha = 1$)	0.433038	4.613957	0.235711	0.612074	0.313807	6.25499	3.62132
Influence ($\alpha = 0.5$)	0.547024	4.733153	0.357063	0.744374	0.414743	6.405727	3.724269
Influence ($\alpha = 0.3$)	0.592619	4.780831	0.405604	0.797294	0.455117	6.466021	3.765448

The top 10 frequent words shown in Table 4 reflect Weibo users′ most attentions in their daily life, especially on the online communications. Specifically, the word of "Livelihood" has occurred 65 518 times in all of the users′ profiles, implying that livelihood is the most important issue in most people′s hearts and minds. The top 10 indicate that the Chinese is peaceful and loves life.

Table 4 Top 10 frequency words in total

Word	Frequency
生活 (ivelihood)	65 518
自己 (oneself)	59 370
爱 (love)	57 317
喜欢 (like)	38 479
关注 (attention)	30 909
世界 (world)	29 169
人生(life)	29 126
快乐 (joy)	27 656
我们 (we)	27 482
幸福 (oneself)	23 417
总数 (oneself)	1 934 428

The analyses above are based on the algorithm as follows: Firstly, to solve the problems of summary field such as short text, sparse features and noisy terms, we develop some novel algorithms on ICTCLAS system[24], for example, symbol recognition, emotion recognition, named entity recognition, language translation, etc. We use these algorithms to extract the most representative keywords of the summary field. Secondly, we use JZSearch platform[25] to build inverted index automatically for all keywords, ensuring the accurate corresponding relationship between keywords and summaries. Thirdly, we can directly get top 10 keywords which represent mostly users′ intention from the inverted index. This step does not need complex computation, so it ensures our system′s efficiency and effectiveness. The last, we use open source tools to implement visualization of keywords (Fig. 14).

All of the data analysis is based on the big data techniques our group has developed, such as big data platform (www.BigdataBBS.com and www.nlpir.org), in the platform, we use our revised patent JZSearch tools[24, 25], which has very powerful computing capability and new generation big data storage management.

Fig. 14 Aword cloud of frequent words

7 Conclusions

With the popularity of microblog in China recent years, SinaWeibo has played more and more important roles in information communications on the Internet. Basic features analysis and numerical features analysis of the Chinese microblog ecosystem have been investigated in this article.

From the analysis and discussed results, several conclusion remarks can be drawn as follows.

First, from basic features distribution analysis, we have obtained the user density distribution of each province in China, in which Beijing reaches top 1 and Macao the 2nd. We also find that microblog is more popular for female users than male, though male users are the driving force. Moreover, usage of young people with high education background are more popular than older ones. In addition, though authenticated users just account for 3% of the total users, they actually pose the remarkable influences in the microblog network.

Second, quantitative features analysis offers an clear result of the distribution of users′ numerical features. Posts-users and followers-users distributions do not fit the normal power law as we knew before. Experiment data show that they both fit the piecewise power law better. On the other hand, by ignoring the noisy points in the beginning of the diagram, the followings-users distribution fits the power law well. Furthermore, we also propose a method to evaluate users′ influence in the microblog network.

User content intention analysis reveals users′ most concerns in their daily life. The word of "livelihood" reaches the top, demonstrating that Weibo users in China care most about their life.

References

[1] Q. Gao, F. Abel, G. J. Houben, Y. Yu. A comparative study of users′ microblogging behavior on Sina Weibo and Twitter. *User Modeling, Adaptation, and Personalization*. Springer, Berlin Heidelberg, vol. 7379, pp. 88–101, 2012.

[2] P. Bao, H. W. Shen, J. Huang, X. Q. Cheng. Popularity prediction in microblogging network: A case study on Sina Weibo. In *Proceedings of the 22nd International Conference on World Wide Web*, Republic and Canton of Geneva, Switzerland, pp. 177–178, 2013.

[3] Sina Weibo. Wikipedia, [Online], Available: http://en.wikipedia.org/wiki/Sina_Weibo.

[4] Z. B. Guo, Z. T. Li, H. Tu, L. Li. Characterizing user behavior in Weibo. In *Proceedings of the 3rd FTRA International Conference on Mobile, Ubiquitous, and Intelligent Computing (MUSIC)*, IEEE, Vancouver, BC, USA, pp. 60–65, 2012.

[5] A. Java, X. D. Song, T. Finin, B. Tseng. Why we twitter: Understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and the 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis*, University of Maryland, San Jose, CA, USA, pp. 56–65, 2007.

[6] A. Mislove, S. Lehmann, Y. Ahn, J. Onnela, J. NielsRosenquist. Understanding the demographics of twitter users. In

Proceedings of the 5th International AAAI Conference on Weblogs and Social Media, AAAI, Menlo Park, pp. 554–557, 2009.

[7] P. Corbett. Facebook demographics and statistics report 2010 [Online], Available: http://www.istrategylabs.com/2010/01/facebook-demographics-and-statistics-report-2010-145-growth-in-1-year.

[8] H. Kwak, C. Lee, H. Park, S. Moon. What is Twitter, a social network or a news media? In *Proceedings of the 19th International Conference on World Wide Web*, ACM, New York, USA, pp. 591–600, 2010.

[9] Q. Yan, L. Wu, L. Zheng. Social network based microblog user behavior analysis. *Physica A: Statistical Mechanics and Its Applications*, vol. 392, no. 7, pp. 1712–1723, 2013.

[10] A. L. Barabási. The origin of bursts and heavy tails in human dynamics. *Nature*, vol. 435, no. 7039, pp. 207–211, 2005.

[11] O. Tsur, A. Rappoport. What′s in a hashtag? Content based prediction of the spread of ideas in microblogging communities. In *Proceedings of the 5th ACM International Conference on Web Search and Data Mining (WSDM′12)*, ACM, Seattle, USA, pp. 643–652, 2012.

[12] J. A. Mangai, V. S. Kumar, S. A. A. Balamurugan. A novel feature selection framework for automatic web page classification. *International Journal of Automation and Computing*, vol. 9, no. 4, pp. 442–448, 2012.

[13] S. A. A. Balamurugan, R. Rajaram. Effective and efficient feature selection for large-scale data using Bayes′ theorem. *International Journal of Automation and Computing*, vol. 6, no. 1, pp. 62–71, 2009.

[14] E. Bakshy, J. Hofman, W. Mason, D. Watts. Everyone′s an influencer: Quantifying influence on twitter. In *Proceedings of the 4th ACM International Conference on Web Search and Data Mining*, ACM, Hong Kong, China, pp. 65–74, 2011.

[15] M. Cha, H. Haddadi, F. Benevenuto, P. K. Gummadi. Measuring user influence in Twitter: The million follower fallacy. In *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media (ICWSM′10)*, AAAI, Washington, USA, pp. 10–17, 2010.

[16] D. Quercia, L. Capra, J. Crowcroft. The social world of Twitter: Topics, geography, and emotions. In *Proceedings of the 6th International AAAI Conference on Weblogs and Social Media (ICWSM′12)*, AAAI, Dublin, Ireland, pp. 298–305, 2012.

[17] D. M. Romero, W. Galuba, S. Asur, B. A. Huberman. Influence and passivity in social media. In *Proceedings of the 20th International Conference Companion on World Wide Web (WWW′11)*, ACM, Hyderabad, India, pp. 113–114, 2011.

[18] Z. Yang, J. Y. Guo, K. K. Cai, J. Tang, J. Z. Li, L. Zhang, Z. Su. Understanding retweeting behaviors in social networks. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, ACM, New York, USA, pp. 1633–1636, 2010.

[19] R. Lee, S. Wakamiya, K. Sumiya. Discovery of unusual regional social activities using geo-tagged microblogs. *World Wide Web*, vol. 14, no. 4, pp. 321–349, 2011.

[20] J. Lingad, S. Karimi, J. Yin. Location extraction from disaster-related microblogs. In *Proceedings of the 22nd International Conference on World Wide Web Companion*, Republic and Canton of Geneva, Switzerland, pp. 1017–1020, 2013.

[21] H. Kwak, H. Chun, S. Moon Fragile. Online relationship: A first look at unfollow dynamics in Twitter. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1091–1100, 2011.

[22] R. Wang, Y. S. Jin. An empirical study on the relationship between the followers′ number and influence of microblogging. In *Proceedings of the 2010 International Conference on E-Business and E-Government (ICEE)*, IEEE, Guangzhou, China, pp. 2014–2017, 2010.

[23] W. L. Chen, S. Y. Cheng, X. He, F. Jiang. Influence rank: An efficient social influence measurement for millions of users in microblog. In *Proceedings of the 2nd International Conference on Cloud and Green Computing (CGC)*, IEEE, Xiangtan, China, pp. 563–570, 2012.

[24] H. P. Zhang, S. L. Wang. Management and Index Method of Perfect Double Array TRIE Tree Dictionary, Patent No. 200510130690. 3. 2005-12-21, China.

[25] H. P. Zhang, Y. L. Huang, C. C. Gong. An Extraction Method and System of Terminology, Patent No. 200710121839.0, China.

Hua-Ping Zhang received his Ph.D. from Institute of Computing Technology, Chinese Academy of Sciences. He is an associate professor of Beijing Institute of Technology (BIT), China. As the team leader, he has been granted over 10 research projects by National Science Foundation, National High Technology Research and Development Program of China (863 Program), National Basic Research Program of China (973 Program). He was awarded the first prize of Qian Wei-Chang Chinese Information Science and Technology Award (top award in the field), president scholarship of Chinese Academy of Sciences (CAS), and Special President Award of Institute of Computing Technology (ICT), CAS.

His research interests include microblog computing, natural language processing, information retrieval and security, big data search and mining.

E-mail: kevinzhang@bit.edu.cn

Rui-Qi Zhang received his B. Sc. degree from Henan University of Science and Technology, China in 2012, and now is a master student in computer science major in natural language processing in Beijing Institute of Technology, China.

His research interests include natural language processing, machine learning and data mining.

E-mail: 843669766@qq.com

Yan-Ping Zhao received her B. Sc. and M. Sc. degrees in mathematics from the Beijing Institute of Technology (BJLG), China in 1982 and 1989, respectively. She was a faculty member at Mathematics Department, Beijing Institute of Technology, and became associated professor in 1989. Now, she is a professor in the Department of Management Science and Engineering, School of Management and Economics, Beijing Institute of Technology.

She has published about 60 refereed journal and conference papers. She received research award from National Science Foundation of China, Committee of Defence Science and Industry of China and many other projects from ministries of China, and won twice the third Awards of Beijing Scientific Progress in 2000 and 2001, Best Paper Award of COINFO Conference in 2009 and a second award from National Education and Engineering Committee in 2008, respectively. She is a member of ACM and CCF.

Her research interests include mathematical statistics, information content security, web mining, e-commerce and e-government.

E-mail: zhaoyp@bit.edu.cn (Corresponding author)

Bao-Jun Ma received his B. Sc. and Ph. D. degrees in management from Tsinghua University, China in 2007 and 2013, respectively. Currently, he is an assistant professor in the Department of Management Science and Engineering in the School of Economics and Management in Beijing University of Posts and Telecommunications, China.

He has published several research papers in international academic journals, such as *Electronic Commerce Research and Applications* and *Journal of Enterprise Information Management*, as well as international conferences. He received several research awards, such as excellent doctoral dissertation in Tsinghua University (2013), Xiao Linshi scholarship for research paper on Chinese economics (2012) and best paper of the Association for Information Management (2012). He is a member of China Association for Information Systems (CNAIS).

His research interests include business analytics and decision-making, information retrieval and search services, as well as policy informatics.

E-mail: mabaojun@bupt.edu.cn