Multi-strategy Based Sina Microblog Data Acquisition for Opinion Mining

Xiao Sun[1], Jia-qi Ye[1], and Fu-ji Ren[1,2]

[1] School of Computer and Information, Hefei University of Technology, Hefei, China
sunx@hfut.edu.cn, lane_3000@163.com, ren@is.tokushima-u.ac.jp
[2] Department of Information Science and Intelligent Systems,
Faculty of Engineering, The University of Tokushima, Tokushima, Japan

Abstract. As an important media for social interactions and information dissemination through the internet, Sina microblog contains emotional state and important opinion of participants. Dealing with microblog data belongs to big data areas, the premise of which is to obtain a large amount of microblog data for further analysis and data mining. For commercial interests as well as security considerations, the access to the data is becoming increasingly difficult and the API Sina microblog officially provided doesn't support large amount of data mining. In this paper, we try to design a platform that is mainly based on the access mechanism of multi-strategy and existing resources to collect data stably from Sina microblog. The results demonstrate that a combination of API and web crawler allows efficient data mining. In such way, sentiment analysis and opinion mining are performed on the data obtained by the multi-strategy method, which proved that the proposed solutions will be allowed to build straightforward application of hot words searching, opinion mining and sentiment analysis.

Keywords: Sina Microblog, Big Data, Data Mining, Web Crawler, Multi Strategy, Opinion Mining, Sentiment Analysis.

1 Introduction

"Behind every microblog user is a living consumer."[3] From testing the water in 2009, to the outbreak of microblog in 2010, and then to triumph in 2011, until Sina microblog apprehension and confusion in 2012, Sina microblog is not prevented from becoming a synonym for web culture. On December 19th, Sina microblog released 2012's annual inventory of web hot words. "London 2012 Olympic Games," "Yan-can" and "PSY" ranked first in three different series, which were totally mentioned in microblog for more than 2.2 billion times [5]. It can be found that Sina microblog with it's over 400 million users have had the power to lead Chinese netizens' opinion. According to Sina fourth quarter and annual financial statements on February 20th, 2013, by the end of December 2012, Sina microblog has attracted more than 500 million registered users, representing a substantial increase of 74%. Daily active users (DAU) reached 46.2 million, accounting for 9.24%. Sina published data, in 2011 and 2012, the development trend of microblog stability, every year the new growth of 200 million users, and the active degree does not reduce [6].

D.-S. Huang et al. (Eds.): ICIC 2014, LNAI 8589, pp. 551–560, 2014.

Recently researching on Sina microblog has gradually become a hot topic. Zhang [4] conducts the research to the network of public opinion based on Sina Microblog, Zhang [2] experiments on Chen Yao's microblog case for example to study microblog propagation mechanism. Wang [10] studies the characteristic and relationship between microblog users. Lian[1] introduces early Sina Microblog scheme of data mining. The acquisition of Sina microblog data mainly two methods: the first one is the use of Sina microblog API access to data; the second method is to use the web crawler microblog webpage access to data, and the two methods are tested. But the original scheme is no longer applicable after the API upgraded, and Sina microblog for its data acquisition have adopted a more stringent restrictions. Comparison of Chen S [9] twitter and Chinese local microblog service focuses on the research on the value of microblog service, Sina microblog as also not full field of study. Researchers deserve attention. Amparo[8] studies on twitter topic sensitivity based on the forwarding path, which has an impact on the user, and the theme content related influence. Fang [11] studies the large number twitter data acquisition scheme based on Twitter List API and Lookup API. LIU [7] tries to dig Chinese microblog interest generated by keyword. More and more researchers begin to concentrate on the emotional content of Microblog distribution, Microblog user's emotional state analysis and other topics.

This paper tries to put forward the adoption of Sina microblog API (OAuth2.0) and the use of web crawler multi-strategy method to parse the weibo.com and weibo.cn on a collaborative program. We try to achieve efficient way to capture Sina microblog data, and so as to obtain valid data to the largest extent in the limited authority. The feasibility of web crawler is tested and the crawling strategy seems perfect. In addition, we use revised SVM Model to training micro-blog emotion analysis model, for trend analysis and passionate tendency for a hot topic in microblog.

2 Data Acquisition and User Database Establishment

2.1 Web Crawler Based Method

The Web2.0 era is arrived, making open platform as the current trend. Sina API is no exception. Developers can submit arguments to the network server and the server issues the requested data to the developer's application. Compared to the way to get data from web crawler, Sina microblog open API interface is more simple and convenient to use. If API calls too frequently, it will lead to temporarily unable to obtain data from the API. All requests will be no return. In order to prevent this, on the one hand by Sina real-time query interfaces under the current access token remaining visits, on the other hand, a program thread of control should be established, with an average frequency of visits.

Microblog data can be requested by calling the API interface and we can analyze them conveniently. But with the coming of the area of big data, the amount of data has become a new standard to measure. So to avoid the data releases too much, the Sina Company updates a series of rules to limit developer catch data. On this occasion, we turned our eyes to the multi-strategy web crawler with web analytic techniques to obtain more information.

The web Sina microblog was divided into PC browser and mobile phone version. Before using the program crawling specified microblog page we need to analyze the

whole login process in detail. Based on the login page source code analysis, we find out that two kinds of algorithms are completely different. On the mobile microblog webpage (http://weibo.cn), the user name password is plaintext transmission when logining. The main parameters of the form are named of password and vk of value. After submit the correct form you will get the cookie to continually catch data. Meanwhile, a long string named gsid is contained in the cookie. If it is put in the request URL, the effect is the same as user being logined to access Sina microblog.

On Sina microblog webpage (http://weibo.com) for PC platform, the process simulation is basically the same. Because of constantly update of Sina microblog, the operation is more complex than microblog for mobile platform, so a detailed algorithm process should be attached. The first step, we need to request the pre-login_URL link and get the parameters such as extract server time, from nonce, pubkey and rsakv values. The username in the submit form should be coded by Base64, the password encryption as RSA, Sina microblog RSA encryption [12] as follows:

① First create rsa public key with two parameters which Sina microblog have given a fixed value when request. These two parameters, the first one is in the first step named pubkey, the second is in the encrypted JavaScript file, it is10001.

②After the encryption process, all the parameters are flexible enough to login. Then request the URL: http://login.Sina.com.cn/sso/login.php?client=ssologin.js (v1.4.4). When the value of the response of the record is 0 indicates successful login. The cookie will be easily got to access other information.

③ In the web page, the Chinese characters are encoded by UTF-8, same as the URL with Chinese characters.

For the above two different ways of logining Sina microblog, acquired data with different features and limitations, which will be compared within the next part.

2.2 Establishment of Microblog User Database

In the experiments and analysis of microblog information, researchers need a large number of complete data package, and for these data, does not require the timeliness is very strong. To the timeliness, whether from the API or through the web crawler, on the premise of all kinds of restrictions, cannot provide timely data very well.

Table 1. User information fields and Microblog content fields

User information fields		Microblog content fields	
uid	username	Meeage_id	Audio_url
screen_name	Sex	Message_text	Video_url
Address	Description	Uid	Repost_count
image_url	attention_num	User_name	Comment_count
fans_num	message_num	Screen_name	Post_time
is_verified	is_daren	image_url	Name_count
verify_info	insert_time	Repost_id	Picture_url
Tag	education_info	Message_url	Platform
career_info	base_info		
create_time	follower_userid		

To a single user, the unique identifier is microblog user's uid, through which we can obtain user's personal information, such as fan list, attention list and even the mood vote in the new version. But for large-scale microblog user's information retrieval, you first need to collect a large number of uid as the entrance to the crawler program. To this question, we artificially screen some microblog uid in good quality, such as hot microblog users. We set the number is 1000 as database. Then we obtain a list of their fans uid through the API to deal with the uid again, expanding the uid library. After looking for a few times, until it reaches the number of threshold setting. From the uid library, we can make use of the crawler to crawl the data. While the efficiency is relatively low, but can get a lot of effective data continually. Compared to the API web crawler has higher flexibility after experiment. In Table 1 above is the user's personal information field and microblog user information fields. Based on these fields we can design data forms to describe and store user's information and microblog data.

Fig.1 below describes the flow chart of user database establishment. The main program can be divided into two parts. One part on the left side is the program which won't stop scanning uid in the database and chose the uid which haven't been used before to grab Sina microblog user's followers to enlarge uid database in order to make sure the part of right side always have work to do. The other part on the right side is a program designed to get user's detail information and the user's whole microblog content data to fill in the two tables above. All the data will gather together to store in the database in case of further research.

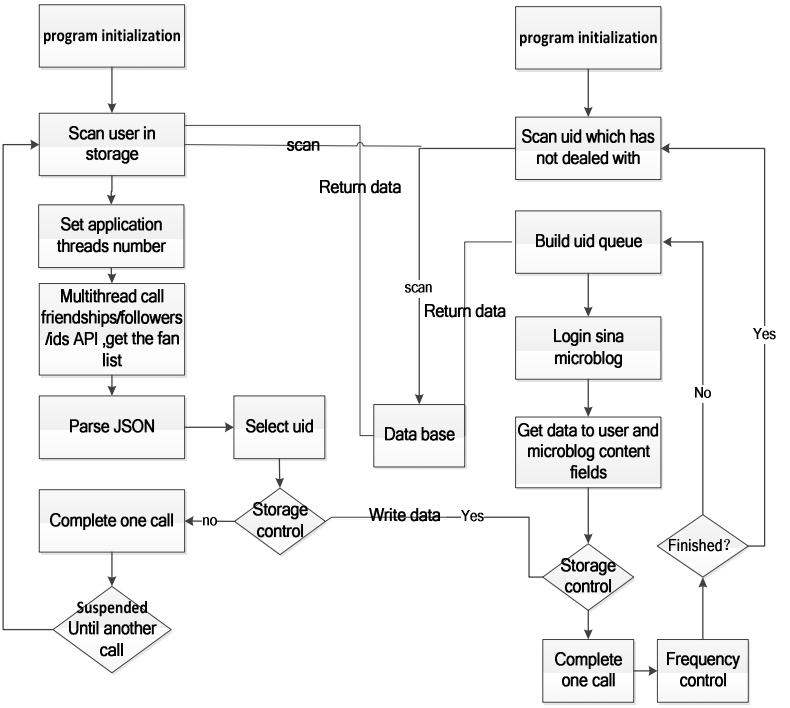

Fig. 1. User database build scheme

2.3 Establishment of Microblog Sentiment Analysis Model

Sentiment analysis of Chinese microblog is mainly divided into text preprocessing, emotional information extraction and emotion classification. Emotional information extraction is divided into emotional words, themes and relationship extraction. The subjective emotional text classification method of microblog is mainly based on semantic dictionary and machine learning [14]. At present, the method based on SVM (support vector machine) of machine learning in Chinese microblog emotion classification is widely used.

Support vector machine is a supervised learning model with associated learning algorithms that analyze data and recognize patterns, employed to classification and regression analysis. In this paper, we choose libSVM package developed by Chih-Jen Lin and revise it as our analysis tool. After observing a lot of microblog with Chinese characters, we find the length of Chinese microblog text is limited to 140 Chinese characters. The text can contain more than one sentence and the emotion in different sentences may be quite discrete. If the whole sentence gives a polarity may affect training with bad effect; more vocabulary and new words on the network appear endlessly. Some of them will become the buzzwords over a period time and get a deeper meaning. Such as "打酱油", "奔疗" etc. There are various forms of expression as well as casual expression. In the process of Chinese microbial emotion analysis researchers need careful consideration on these factors.

Xie [15] and other researchers experiment on microblog data forms for rules of emoticons method and emotional dictionary method. Moreover they applied the hierarchical structure of the SVM multi-strategy method. The experimental results show that compared with the former two methods of rules, the method based on SVM is best. They further analyzed the characteristics in detail. In this paper, we will use the method of multiple SVM classifiers training emotional classification model to judge the microblog emotional tendency. The emotional dictionary is HowNet and emotional words library from Dalian University of Technology. Training the SVM classifier process is as follows:

(1) Depending on the method of subjective and objective classification feature description [15]. Extract microblog content classification of subjective and objective characteristics from the data forms. Then start training objective and subjective SVM classifier.

(2) Extract polarity classification features from training corpus clauses, then using chi-square statistics to choice feature words. The chi-square formula is presented in formula (1):

$$\delta^2(w,e) = \frac{(\mathrm{AD}-\mathrm{BC})^2}{(\mathrm{A}+\mathrm{B})(\mathrm{C}+\mathrm{D})} \tag{1}$$

W is candidate words; e means the emotional tendency where w is it; A means the number of microblog which contains word w and belong tendency e; B means the number of microblog which contains word w but doesn't belong tendency e; C means the number of microblog which doesn't contain word w but belongs tendency e; D means the number of microblog which neither contains word *w* nor belongs tendency *e*. After the feature words have been generated, we use the TF-IDF to calculate the

weight of feature words. Then according to the positive and negative tag we can train SVM classifier.

3 Experiment and Results

The experiment platform is SONY CS36 notebook, Intel core P8700 dual-core, ECLIPSE EE program platform, 4 GB of memory, WINDOWS 7 platform, the program is developed by Java, with access networks of 10 MB sharing education network.

3.1 Comparison of File Size of Query Results

In this paper, based on the Sina microblog API and web crawler, we design total of three kinds of Sina microblog data acquisition scheme. In the test, we first manually choose 100 Sina microblog users whose fans is more than 5000 people, in order to make sure all of the user's data are available through the API, and web crawler. Sina microblog API can only return 5000 fans uid information, and at weibo.com, the uid information can query to only 1000 fans. In http://weibo.cn and can query to the 5000 fans as well as API.

Table 2. The comparison of the file size of query results in user fans

Request	total fans	total files	file size
JSON	500000	100	4.95MB
weibo.cn	500000	500	558MB
weibo.com	100000	100	442MB

Table 3 shows the size of queried uid whose fans is over 100 users. During the test, JSON is built on the API query results. Compared through the API to get the user fans, we can find web crawler gets the data more vast. On the premise of getting the same amount the number of fans, the file size is almost one hundred times more. This is because the web crawler retained a large number of useless information in the text, including the HTML source, JavaScript tags. So, the efficiency through the way of web crawler to parse user state is less than the approach to get user state by using the API method.

3.2 Access Request Number Limitation Test

The request limitation of the API can be found directly from the open platform of Sina microblog and also can find that from the API documentation. Access restrictions on weibo.com and weibo.cn, we cannot have a clear data on the Internet. Therefore, we design crawlers to maximum to the limit of single account's request number. The experimental results are shown in Table 4 below. We can find that, for web microblog (weibo.com), the user profiles and user access is not restricted, so it can be used as building microblog user database. And the mobile phone microblog (http://weibo.cn) can also be used as a data access or a secondary data source. For the application of crawler on keywords search, Sina microblog have the corresponding limit. Looked at the API it is not provided, it just has been able to get part of the data. For this limit, we need to establish an account pool so that we can gain a large number of data.

Table 3. Request times on crawler limit

crawler way	weibo.com	weibo.cn
Data access	unlimited	unlimited
Keyword	40times/h	500times/h
State of the user access	unlimited	500 times/h

3.3 Search and Analysis of Hot Keywords from Microblog

Researchers can perform the corresponding research after having the corresponding source of data. The following is keywords search and analysis, for example. In the first half of 2013, the food safety issue often appears on microblog. Our experiment was fixed on the dead pig events in the Shanghai area. We selected a list of keywords and search the keywords by province. Acquisition time is taken from March 3, 2013 to April 5, 2013. Because Sina microblog API does not provide the corresponding interface, therefore the experiment first registered a series of microblog account, then get microblog content of 34 provinces and regions through the crawler. Finally we have collected a total of more than 68000 microblog content. The microblog provinces distribution of the data is shown in Fig.2 and Fig.3 below.

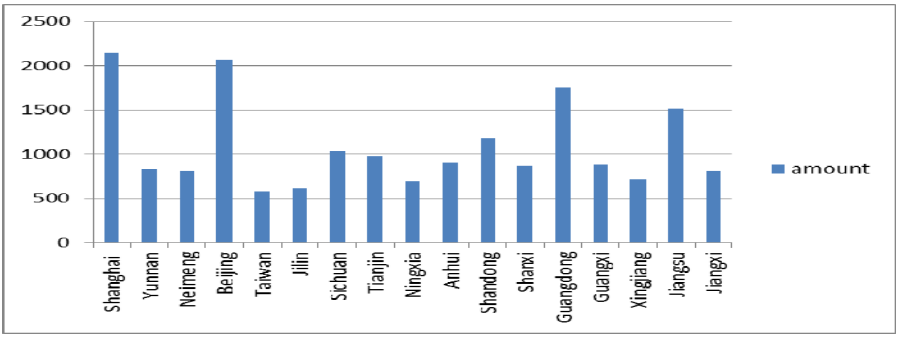

Fig. 2. Microblog number provinces distributio part 1

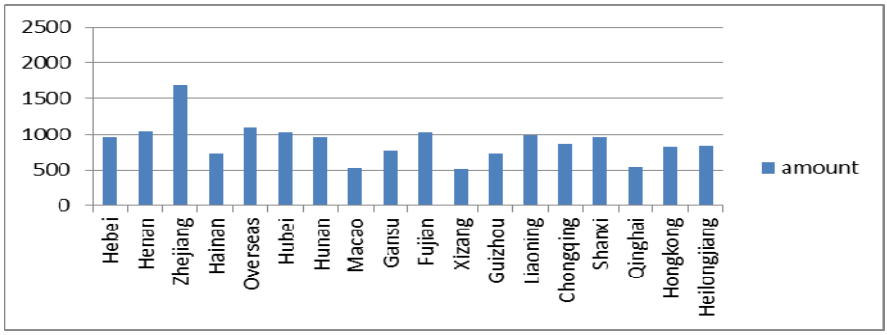

Fig. 3. Microblog number provinces distributio part 2

From Fig.2 and Fig.3, of the dead pig event, the number of microblog in Shanghai, Beijing, Guangdong, Jiangsu, and Zhejiang is more than other province. Shanghai, Jiangsu, Zhejiang, Guangdong are the areas where dead pig event affected. And Beijing area's number is large because the government department is located in the Beijing area, which mainly releases news about dead pig incident investigation process and results by means of the official microblog. The number of microblog on other provinces is in a low degree. This visible indicates that the whole country is highly concerned about food safety.

In addition to microblog province distribution, microblog released time are also used to statistics. As showed in Fig.4. It can be found in the line chart that the overall trend of microblog release quantity takes on the form of two peaks. Around March 10, microblog release quantity began to increase, to March 14, began to decline and maintain in a certain number, and then around March 28 microblog release quantities began to increase again, then to April 1 began to decline. Microblog quantity changes around an event topic have close relationship with the event processing of the incident.

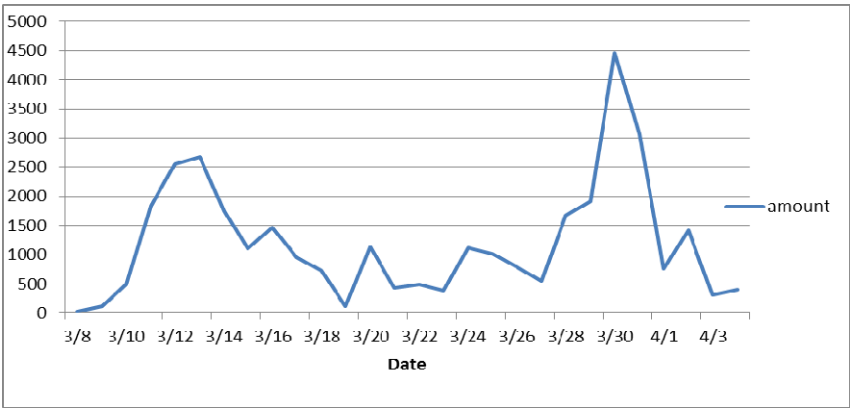

Fig. 4. Microblog release time distribution

Microblog is the expression of public opinion platform. Public emotional changes when events with incident investigation going through. In the microblog event sentiment analysis, we can estimate the tendency to the view of public opinion. To figure out these emotions, we first prepared manually labeled training data, which was comprised of 1000 microblog from the datasheet randomly selected. We manually annotate them as positive examples and negative ones. We built a classifier based on support vector machine. We use libSVM with a polynomial kernel. We use the features above and 10-fold cross-validation to find the proper parameters c and g. The main steps follow the steps described in section 2.4 above.

Topic microblog processing is as follows:

1. Divide the sentences.
2. Predict the microblog subjective and objective tendency by SVM model

3. To the subjective microblog use polarity SVM model to predict their polarity tendency, and the according to the number of positive and negative numbers to classify microblog according to the formula (2):

$$\begin{cases} positive\ (positive\ sentence\ number > negative\ sentence\ number) \\ negative\ (positive\ sentence\ number < negative\ sentence\ number) \\ neutral\ \ (positive\ sentence\ number = negative\ sentence\ number) \end{cases} \tag{2}$$

The experimental results are shown in Fig.5. From the results, microblog negative emotion change trend similar to microblog release time of the trend. When microblog release quantity increases the negative emotion increases; positive microblog emotional change trend was smoothly from the beginning to fluctuations.

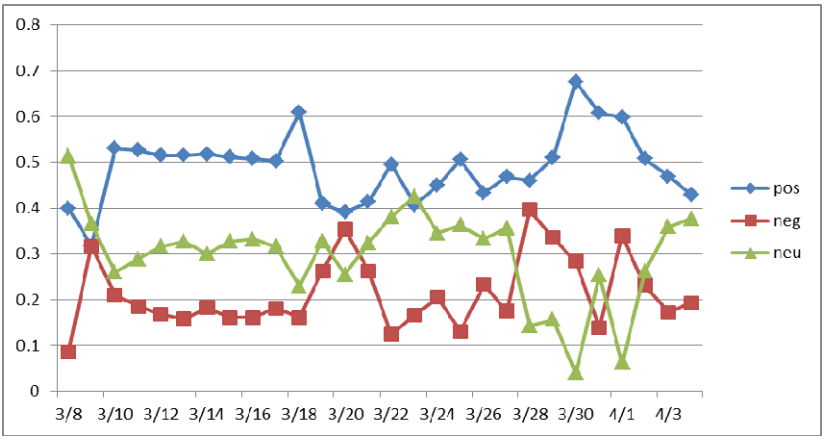

Fig. 5. Microblog emotional distribution

4 Conclusions

With the popularity of microblog in social interaction, microblog text analysis research has also become a research hotspot nowadays. In order to facilitate the microblog text analysis experiment, this paper introduces a design of data mining to microblog: a small number of user's uid for the entrance, smoothly get plenty of microblog data.

This paper first analyzes the Sina official API, the usage and limitations are described, and then analysis of two kinds of webpage microblog login process. After that, we designed the storage of users' data and completed the process of data mining. Finally the experiments show that, through the combination of API and multi-strategy web crawler can efficiently obtain microblog data. At present, the method also has a few shortcomings; we will continually develop it from the following aspects, for further research and exploration.

First of all, in Sina microblog data mining scheme, there has many customized mining scheme, in order to adapt to different data requirements. We need to solve the efficiency and low cost problem to get the needed data for research.

Secondly, to microblog data analysis, large-scale data is needed. Data with increased time mean the storage scheme should redesign and how to better organize these microblog data need further consideration. In addition, relations of users in microblog data are very valuable, on how to effectively deal with it also needs further study.

Acknowledgment. This research is supported by National Natural Science Funds for Distinguished Young Scholar(No.61203315) and 863 National Advanced Technology Research Program of China (NO. 2012AA011103).

References

1. Lian, J., Zhou, X., Cao, W., Liu, Y.: SINA microb log d a ta ret ri ev al. Journal of Tsinghua Un iversity(Sci & Tech) 20, 11(5),1(10),1300–1305
2. Zhang, Y.X.: Sina microblog dissemina-tion mechanism research. Southwest University (2011)
3. Xie, A.P.: Look from the network market-ing tools using vancl development. Discovering Value (4), 21 (2011)
4. Zhang, L.L.: Sina microblog public opinion analysis and research. East China Normal University (2011)
5. The 21st century economic report, to find the new world in 2013 on sina microblog (2013), http://jinhua.house.sina.com.cn/news/2013-02-23/10292163302.shtml,2013-02-23
6. Sina technology, Sina microblog 2012 annu-al inventory (2012), http://tech.sina.com.cn/i/2012-12-19/13447902817.shtml,2012-12-19
7. Liu, Z., Chen, X., Ssun, M.: Mining the in-terests of Chinese microbloggers via keyword ex-traction. Frontiers of Computer Science 6(1), 76–87 (2012)
8. Cano, A.E., Mazumdar, S., Ciravegna, F.: Social influence analysis in microblogging platforms–A topic-sensitive based approach. Semantic Web.
9. Chen, S., Zhang, H., Lin, M., et al.: Com-parision of microblogging service between Sina Weibo and Twitter. In: 2011 International Conference on Computer Science and Network Technology (ICCSNT), vol. 4, pp. 2259–2263. IEEE (2011)
10. Wang, X.G.: Empirical analysis on behavior characteristics and relation characteristics of microblog users – take ”sina microblog ” for ex-ample. Library and Information Service 54(14), 66–70 (2010) (in chinese)
11. Fang, W.W., Li, J.Y., Liu, Y.: Re-search on twitter data collection. Journal of Shandong University(Natural Science) 47(5) (2012)
12. Microblog login with program (Python and RSA encode algorithm) (May 5, 2013), http://www.2cto.com/kf/201303/192970.html (in chinese)
13. Microblog open plate-form, http://open.weibo.com/,2013-3-15
14. Zhou, S.C., Zhai, W.T., Shi, Y.: Overview on sentiment analysis of chinese microblog. Computer Application and Software 30(3), 161–164 (2013)
15. Xie, L.X., Zhou, M., Sun, M.S.: Hierarchical structure based hybrid approach to sentiment analysis of chinese microblog and its feature extraction. Journal of Chinese Information Processing 26(1), 73–83 (2012)
16. Sina news report about Shanghai dead pig event (April 2013), http://roll.news.sina.com.cn/s_sizhu_all/index.shtml