Research on the Grabbing and Application of the Big Data of Sina Microblog Based on Multi-Strategy

Tianding Zhang[1, a]

[1] Major in Computer Science, College of Liberal Arts, University of Minnesota, 55455, America

[a]email

Keywords: Sina microblog API, Data mining, Multi-strategy

Abstract. Micro-blog has become a channel to obtain the information. The methods of obtaining microblog data are becoming less. With the upgrading and renewal of Sina API, the channel accessing the data sources is limited. From the current situation of data acquisition, the author tries to analyze the best way to get the micro-blog user data at present. There are two common kinds of methods. One directly acquires the user's data and information through the analysis of micro-blog page; the other obtain data from the official micro-blog API. These two methods are very difficult to operate, and it is not easy to grab the data completely.

Development History of Sina Microblog

Micro-blog has established for 8 years. From the political issue in the early stage of the discussion gradually to the development of interest in social networks, social information content gradually gave way to star entertainment culture and other content. Big V had about micro-blog also gradually gave way and more vertical field of small V, you can clearly feel. Whenever there is a hot event when micro-blog will soon appear corresponding topic, and the sound. This kind of one-way open mode of communication is similar software cannot do, in fact, caught the wave of video and live. In the second quarter of 2010, micro-blog video playback volume increased by 235% compared with the previous quarter, live screenings over 10 million, 116 times higher than the previous quarter. The text in the form of live video and pictures, micro-blog network red provides an excellent stage to create content, attract attention, commercial realization, network economy is doing the so-called red operation, and micro-blog declined significantly, the strategy of active users can maintain stable growth. Although you are in the first-tier cities, you feel less and less using micro-blog, but for the 234 line of Internet users, micro-blog is still the most important channel for access to information. Data show that the fastest growth of micro-blog MC is second tier cities, followed by the three tier cities and four tier cities. Wall Street is also optimistic about micro-blog, by 2020, online advertising scale will account for more than 80% of China's advertising market. Micro-blog is one of the most dominant in the areas of the company, which is expected to maintain more than 20% growth rate in 2016 and 2017.

Summary of Data Grabbing Methods

The user state, users and fans and other information have become very large and valuable data, but these data are only micro-blog's own official processor inside, outside want to get these data and information is very difficult. The web crawler was born in this situation and how to improve the operation of the crawler strategy become a problem in data grabbing of today's micro-blog.

Data Grabbing Based on API. API is the abbreviation of Application Programming Interface. Specifically, is a site of accumulation and the change of the data, and these data if sorted out, not only time-consuming, also a lot of space, there is a danger just finishing well. Most people need the expired data, in fact, are just a small part, timeliness requirements may be very strong therefore. Organize the storage and provide a download to the public is not cost-effective. But if not in some way to open the data, will also face numerous reptiles' harassment. This will bring a lot of trouble to compromise. The normal operation of the website, the website is to provide a channel. When you need a portion of data although, none of the available data sets, but only need to use this channel, you

want to describe the data, and then through the website after the audit, that can give You will immediately send the data you requested. The two sides clearly come to the satisfaction of all. So, in the future when you find data. It may also wish to look at the target site, which is provided by API to avoid doing useless work.

To Use OAuth2.0 to Call API of Sina Microblog. Sina micro-blog SDK provides Oauth2.0 information collector. 0 authorized certification, at the same time it also has a strong independent SSO login function, cannot go through the official verification can enter the information interface, and it also has micro-blog copy share function, can use the software directly to the need for information interception down we use micro-blog Android and Sina SDK. There are three types of authorization, but the three landing must agree to Oauth2. 0 protocol can enter the system, the main content of this agreement is paid through the way we can provide the user with a login token, the token is equivalent to our login account and password, and this token can also be used to store some data, these data are set themselves according to the needs of users, it can we allow the user to access the data stored in specific. This will greatly reduce the landing steps of OAUTH, give the user access to micro-blog authorized user information stored in another server, without specific sharing can get information they need.

Comparison between Sina Microblog and SDK. The simple definition of API is to provide general service, may the smaller particles, because of the need to consider the reuse. Service SDK is a simple consumer service set, a general SDK is the integration of multiple API integrated with client session attributes, reaching more consumer business logic, such as the application of SDK in charge only API call, security control, cannot replace the two times to confirm the interface. Therefore, the difference is mainly reflected in the development process of service use, SDK may be simple, native API may be complex, but SDK itself may be a problem, such as the implementation of SDK some basic abilities (such as network) is not perfect, or SDK the interface with your UI style is not consistent.

Information Grabbing of Microblog Based on Multi-Strategy

For the information acquisition of Sina micro-blog, we can directly call the API interface, which gives the data access and transfer provides a new method of Sina. Micro-blog uses many REDIS technology. The difference in MEMCACHED, REDIS periodically updated data is written to disk or to modify write operations additional REDIS book files, and on this basis to achieve master-slave synchronization. This need to set up RESIS server. However, in the aspect of data access and call for API interface in a lot of time is not to allow the user to access the advanced information. So, this paper will introduce the next web crawler and Sina micro-blog landing procedures simulation algorithm.

Sina Microblog Simulation of Landing Program Algorithm Process. Sina micro-blog simulated landing program algorithm with Facebook Top Story algorithm is similar, it pays special attention to the factors considered intelligent sorting algorithm and specific weight, it needs to determine a dynamic score by many factors, the final evaluation index is adjusted to a desired. These factors can be roughly divided into three categories: (1) timeliness, in the intelligent sorting algorithm, timeliness is no longer like before, is decided. The only factor in micro-blog ranking but micro-blog is active users, online are very high, timeliness should be still very important factors. Of course, this factor may adjust the frequency according to the different landing. The user (2) the quality of content, for a micro-blog, to determine the content of quality factor is divided into three parts: the weight, whether as a member of the creators of the certification So, depending on the quality of the content determination method for Sina micro-blog may also include other sources. But not all the above factors, may choose one of the most important factors involved in sorting. (3) personalized, personalized is through a variety of data to calculate what you love, love the topic areas and micro-blog account at the same time, your account will get higher scores for Sina. Micro-blog. Micro-blog is currently the intelligent sorting method with user activity to reduce the corresponding measures taken.

Two Completely Different Landing Forms. Crawler is a simulation using the code browser to access or download a browser page and based on the simulation tool. The structure of HTML filter to obtain the required information in R. We usually use the RCURL package to achieve the access function, using the XML package by HTML tree structure. The first step of screening information extraction, simulation of browser behavior, if you want to use the R language the simulation of browser behavior, we must disguise. When we use the header in the browser to access a web page, the browser will send some instructions as a web server. The second step is to simulate the page. This process in fact is to specify the "temporary download. The third step is to use HTML structure. Finishing the function analysis in XML just get the variable (which is stored in the HTML web crawling code) to generate standard HTML tree structure and assigned to the variable Then. In the XML of the variable to carry out various operations. The fourth step, the node location, if we want to get the two labels. It can be used for locating to write your own code, through the above four steps, and we can get the data we want.

To Use OAuth2.0 to Call API of Sina Microblog. Sina micro-blog SDK provides Oauth2.0 information collector authorized certification, at the same time it also has a strong independent SSO login function, cannot go through the official verification can enter the information interface, and it also has micro-blog copy share function, can use the software directly to the need for information interception down we use micro-blog Android and Sina SDK. There are three types of authorization, but the three landing must agree to Oauth2. 0 protocol can enter the system, the main content of this agreement is paid through the way we can provide the user with a login token, the token is equivalent to our login account and password, and this token can also be used to store some data, these data are set themselves according to the needs of users, it can we allow the user to access the data stored in specific. This will greatly reduce the landing steps of OAUTH, give the user access to micro-blog authorized user information stored in another server, without specific sharing can get information they need.

Database Establishment of Microblog Users. First of all, we need to design a program, then this program is not according to query the user query to the unit or not registered in the UID, then we based on the query to the data of the number of threads application type setting, and construct the UID formation task allocation, then we use multi thread calls API interface, access to the user list of fans micro-blog, Sina landing simulation, then we parse the JSON object file, extract the user according to the UID screen to grab the user related information, to build and fill Sina micro-blog user model, then this model storage control, this completes the API call, but also to complete the micro-blog crawler call, but we must strictly limit the frequency we visit, waiting for the next visit..

Analysis Model Establishment of Microblog Emotion. Chinese emotional micro-blog text analysis is to analyze a sentence is subjective or objective description of subjective description, micro-blog text sentiment classification method is mainly based on semantic dictionary machine program to judge. At present, based on SVM (support vector machine, referred to as SVM) in micro-blog Chinese polarity judgment is widely used study methods of machine learning.

First, to determine a word is positive or negative, is subjective or objective, this step mainly rely on the dictionary. Chinese, positive and negative judgments have many dictionaries, but resource dictionary resource quality is not high, not careful. In addition, the lack of subjective and objective judgment and dictionary statement. The next is to identify a sentence is positive or negative, it is subjective or objective. There will be many simple dictionaries. Directly to match a sentence what dictionary words, then we can calculate the summary sentence. But the sentiment scores due to the different fields have different emotional words, such as "blue screen", this word generally does not appear in the emotional dictionary, but the words clearly expressed dissatisfaction. So, we must according to the specific field to construct a specific emotion dictionary. If not So much trouble, you can use supervised machine learning method. A bunch of comments into a training algorithm which, after training the classifier can put into positive and negative comments, subjective and objective. At last, the emotion mining upgrade to opinion mining. For example, analysis for a product, which requires the based on the analysis of emotion, to dig out the nature of the product, and then analysis

the corresponding attribute of emotion. After analyzing every review all the attributes of emotion, you can put together, the formation of consumer evaluation of a product of each part.

At the same time, the algorithm is very representative, the specific contents are as follows: the first step: read the comment data. The second step up the search for clauses comments: emotional words, recorded positive or negative, and position. The third step: to find the degree of emotional words before the word, find it. Stop searching set weights for word level the fourth step, multiplied by the emotional value. To emotional words before finding negative words, find all the negative words, if the number is odd, multiplied by -1, even if, multiplied by 1. Fifth step: to determine whether there is an exclamation point at the end of the clause, there is mark go looking for emotional words, there is a corresponding emotion value +2. The sixth step: calculate a comment clause all the emotional value, with an array of records. The seventh step: calculate and record all the comments. The eighth step: positive emotion value calculation of each comment clause by love the sense of mean, mean positive emotion negative emotion, negative emotion variance, variance. According to this algorithm, we can design the model.

Search and Analysis of Hot Key Words in Sina Microblog. The first half of 2013, the food safety issues in micro-blog often appear in this experiment. Keywords dead event in Shanghai area as an example, all provinces and regions of micro-blog keyword search. The acquisition time is from March 3, 2013 to April 5, 2013. Due to the Sina micro-blog API does not provide the corresponding interface, so the experiment first registered a series of micro-blog account to access micro-blog the contents of the 34 provinces and regions through the crawler, sampling statistical analysis. A total of more than 68000 micro-blog acquisition items.

Search and Analysis of Hot Key Words in Sina Microblog

Micro-blog hits around a theme of the event changes, with the development of this event has a close relation to the event. Micro-blog sentiment orientation analysis, can be inferred. Opinion for micro-blog's topic as follows:

(1) The first is to separate the clauses for micro-blog.

(2) To use subjective and objective clauses to predict the SVM models

(3) The subjective micro-blog SVM polarity model is used to predict the emotional polarity, and then the micro-blog is classified according to the number of positive and negative clauses

Positive emotion (positive clause number > negative clause number);

Negative emotion (positive clause number < negative clause number);

Neutral emotion (positive clause number = negative clause number).

Microblog negative emotion changes with the time microblogging released a similar trend, increased with the amount of micro blog; microblog positive emotional changing tendency from the first wave steady turning; neutral emotion is microblog and negative emotions like fluctuations.

Conclusions

Sina micro-blog follows and gains twitter information flow model, and successfully spreads in China, greatly promoting the opening of Chinese public opinion ecology. It can be said that micro-blog is an exciting product, and it is not an empty talk to change China. Now micro-blog has entered the mature period of the community, and a normalized product has faded out of the original aura. It is very commendable.

This paper introduces the development process of Sina micro-blog, and then explores the methods of acquiring the user’s data of Sina microblog. We also explain the issues relating to the technology program of micro-blog and the Chinese judgement of microblog in detail. This paper is only a simple introduction, hoping to be able to cause the attention of the related parties to the data acquisition. At the same time, the author also hopes to receive the valuable comments and suggestions from the professionals.

References

[1] Sun Xiao, Ye Jiaqi, Tang Chenyi, et al. Method of Sina microblogging big data grabbing based on multi-strategy and its application [J]. Journal of Hefei University of Technology (Natural Science), 2014, 37(10): 1210-1215.

[2] Zhu Yunpeng, Feng Feng, Chen Jiangning. Chinese microblog data collecting method based on multiple hybrid strategies [J]. Computer Engineering and Design, 2013, 34(11): 3835-3839.

[3] Yang Fei, Jiang Nan, Li Xiang, et al. Research on the Method of Microblog Location Data Acquisition Based on Multi-Strategy [J]. Journal of Geomatics Science and Technology, 2016, 33(2): 201-207.

[4] Xiang Yu, Guo Yunlong, Xu Xiao, et al. Entity Words Disambiguation and Entity Linking with Multi-Strategy in Chinese Microblogs [J]. Computer Applications and Software, 2016, 33(8): 12-17+61.