An Advanced Data Capture Method based on Sina Weibo

TianSheng XU, Xun YANG, HongChen ZHANG, Jiong ZHANG
School of Information Capital University of Economics and Business, Beijing, China

ABSTRACT: Owning more than 50 million daily active users, Sina Weibo offered a large amount of direct or indirect raw data and is selected as the data source of this study. In this paper, A new scheme on data capture is developed, which is based on the specified keywords provided by demand. It mainly focuses on how to use web crawlers to grab data during the specified time and verified by Sina Weibo data.

KEYWORD: Sina Weibo; Web crawlers; data grabbing; keyword searching

1 INTRODUCTION

In recent years, as a way of online communicating, the social media has been developed very fast and the most prominent is the social network, Such as FaceBook, Twitter, MySpace, Sina Weibo, RenRen, KaiXin, Tecent (PengYou Web, Weibo). Huge amounts of data are generated by these social networks every day and have not been deeply processed yet.

2 THE RESEARCH VALUE OF SINA WEIBO

By using Twitter data, Bernardo A. Huberman, a director of social computing study department from HP Labs, HP research department, together with his colleague Sitaram Asur, used tweets frequency model they have created to predict American movie box office revenue and the result they predicted was more accurate than that from forecasting department which based on film market. By using Twitter data, four scholars from Singapore Management University predicted 2011 Singapore election results and 2012 America election results and their predicting result was more accurate. By using Twitter Tweets to predict the Dow Jones index Scholar Johan Bollen from Indiana University Bloomington designed and implemented the algorithm. This algorithm's forecasted accuracy on Dow Jones index's daily closing price as high as 87.6%.

Sina Weibo also has great research value. This article will focus on the first critical step on Sina Weibo data mining, data capture. After analyzing the pros and cons of different current capture methods, we choosed a fetching strategy which suited for time series analysis. According to the specified keywords We designed fetching programs and eventually they achieved our goals that captured data during specified period of time..

3 THE COMPARISON OF DIFFERENT DATA CAPTURE METHODS

The data capture methods on Sina Weibo mainly include Sina Weibo API and web crawler, or a combination of both applications.

3.1 *Sina Weibo API*

Based on the massive amount of users on Sina Weibo and its powerful ability on propagating its data, Sina Weibo open platform is such a kind of open platform that it accesses third-party partner services, not only providing users rich applications but also improving its service. Accessing your service to Weibo platform can help to promote their products and increase website / application traffic, expanding new users and obtain benefits.

To facilitate users develop kinds of applications themselves, Sina Weibo provides a relatively consummate application programming interface. Among them, relative interface on Weibo data capture is search/topics, whose function is to search Weibo data without giving topics. This definition directly determined that it doesn't have that function which searches according to specific keywords. Meanwhile, this function clearly described that the interface can only return the latest 200 results.

After analysing each interface of Sina Weibo, we find that none of them provide a function searching Weibo by keywords or by specified time period. In addition, Sina Weibo API also sets up some restrictions like accessing frequency. As a result, in this article we do not use Sina Weibo API to crawl data.

3.2 *Web crawler*

Web crawler (also known as web spider, web robot, website chaser) is a program or script. It can automatically crawl web data according to some certain rules. Sina Weibo is mainly produced in the mobile Internet environment,so that it is particularly suitable to use web crawler to capture the relevant data needed for study. Meanwhile, for those advanced search functions in this article, web crawler need to solve such problems like user login, access restrictions, data extraction and other issues.

3.2.1 *Premise of advanced search*

Analysis shows that if we only search the keywords without setting time parameters, Sina Weibo will allow users search relevant keywords without logging in. However, the purpose of this paper is not only to specify the keywords, but also requires a specified time period to capture data. That requires us to use advanced searching function of Sina Weibo. But Sina Weibo regulates that we must log in if we want to use the advanced searching function.

3.2.2 *Visit restrictions*

Sina Weibo sets up IP visiting restrictions on times, see Table 1 Sina Weibo access frequency.

Table 1 Sina Weibo access frequency description

Authorization Level	Visits Restrictions
Testing Authorization	1000 times per hour
General Authorization	10000 times per hour
Intermediate Authorization	20000 times per hour
Advanced Authorization	30000 times per hour
Cooperation Authorization	40000 times per hour

All those authorizations except testing authorization are for application developers. General web crawler corresponds general authorization, that same IP requests can not exceed 1000 times per hours. This section requires that the design of web crawlers in this article must set up corresponding function against visit restrictions. The aim is to avoid our requests being rejected.

3.2.3 *Data extraction*

In this paper, the data which captured by web crawler will be saved as HTML text. Meanwhile, this file contains a large number of redundant HTML tags, hyperlinks and other useless data. We need to design and implement a function which can extract specific data and remove useless information and store relevant Weibo data.

This paper focuses on these three issues to solve. As well as we solve them we also obtain the function which can capture data from Sina Weibo during specified time period and specified keywords.

4 DESIGN OF DATA CAPTURE PROGRAM AND ITS IMPLEMENTATION

After a comprehensive analysis of the advantages and disadvantages of Sina Weibo API and web crawler, as well as the special search requirements (specified time period, specified keywords), we take web crawler to fetch data in this paper.

5 SIMULATING LANDING ON PYTHON

5.1 *Simulating landing on Python, specific processes:*

5.1.1 *Step A*

Construct request. Encrypt usernames and passwords. Construct request string.

5.1.2 *Step B*

Get login Cookie. Send the request string in step a, save the Cookie returned from servers.

5.2 *Use thread to control visit times*

By introduction of thread mechanism we mainly ensure that the number of visits times per hour is less than 1000:

a) Construct parameter requesting string which is corresponding to advanced search.

b) Send requests and save the search results.

c) Programs wait for a certain time under the thread control.

d) Determine whether the crawl is completed by the termination condition, if not, loop to step a.

5.3 *Data extraction*

Initial data we captured is saved as HTML formatted text. We need to remove redundant data like HTML tags and useless hyperlinks, and ultimately save these data to a MySQL database.For this reason, we mainly use BeautifulSoup.

5.4 *BeautifulSoup*

BeautifulSoup is particularly suitable on parsing Internet web page. It has very perfect function on XML and HTML and also provides traversal functions according to class, style and other attributes.

In those HTML texts we captured, all Weibo contents are included in a tag named dl whose class attribute is feed_list. For further analysis the contents of each tag, we set keyword as 'Shanghai Composite Index'. Analysing the document structure, we select the corresponding function of BeautifulSoup, mainly use its find_all () method. Method name: find_all (name, attrs, recursive, text, limit, ** kwargs) The meaning and function of each parameter are given a detailed explanation on BeautifulSoup official document.

Function final_all() filters out specific label according to the parameter list by traversing all the document labels. In this paper, we mainly use the function which can filter off specific label according to CSS class. The specific method is: find_all (“dl”, class_ = “feed_list”). We can filter off all the dl labels whose class attribute is feed_list, while use regular expressions to remove excess HTML tags, so that we can only save Weibo's time and contents. Disjunction process:

a) Reading the HTML file.

b) Construct BeautifulSoup object.

c) Get published time and Weibo content according to the label style.

d) Saved to database.

e) Determine whether the disjunction is completed by the termination condition, if not, loop to step a.

In solving the key issues of web crawler application, this paper presents a Sina Weibo data capture process according to specific time period and specific keyword. As Figure 1 Capture Process shows:

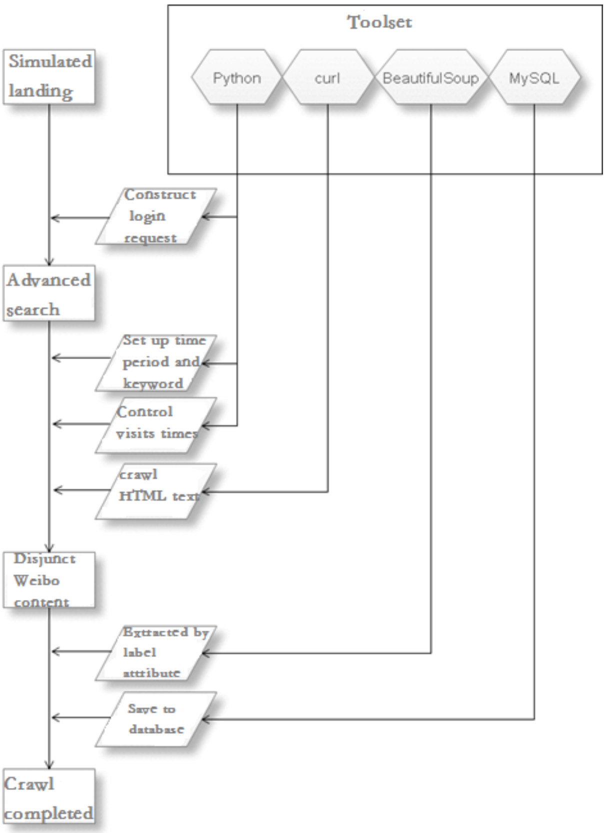

Figure 1 Capture Process

6 CONCLUSION AND ANALYSIS

This article uses Python to write web crawler, combined with curl, BeautifulSoup, MySQL and other tools to achieve the purpose that capture Sina Weibo data from specific time period and specific keyword. Meanwhile, in order to verify the effect of crawler for keyword "Shanghai Composite", we fetch three years' Weibo data from 2010-01-01 to 2012-12-31, running on PC. Table 2 web crawler operating environment:

Table 2 web crawler operating environment

Computer Type	ThinkPad Laptop
Frequency	2.3GHz
CPU	Intel i3
RAM	2G
Operating System	Windows7
Python Version	Python 3.3.0
BeautifulSoup Version	BeautifulSoup 4.0
MySQL Version	MySQL 5.6
Curl Version	curl 7.33.0

Operating result, see Table 3 Test Results:

Table 3 Test Results:

Keyword	Shanghai Composite
Time Period	From 2010-01-01 to 2012-12-31
Crawl Time	2 hours and 24 minutes
Crawl Weibo Records	57722 records

The conclusion on the research of Sina Weibo data grabbing in this paper can solve problems like simulating landing, accessing frequency limit, data extraction in data capture. The scheme proposed can also achieve data crawling by specific keyword and specific period of time. It establishes the data foundation on social network for data mining and is helpful for researchers to predict future by data of Sina Weibo or do some other related studies. According to the test results, it seems that the operating efficiency of the program need to be improved. Multiple IP strategy combining with the distributed system will be studied in our future study.

ACKNOWLEDGMENT

This research was supported by Funding Project for Research Level's Improvement by CUEB and Research Project of CUEB (2014XJG022). It was also supported by Level of Scientific Research Improving Project of Beijing Municipal Education Commission 2014 and Beijing philosophy and social science project (13SHB015), and the Funding Project for Research Level's Improvement by BMEC.

REFERENCES

[1] Johan Bollen, Huina Mao, Xiao-Jun Zeng, Twitter mood predicts the stock market, NSF Grant BCS, 2010, 10: 1~8

[2] Andranik Tumasjan, Timm O. Sprenger, Philipp G. Sandner, Isabell M. Welpe, Predicting elections with Twitter: What 140 characters reveal about political sentiment, International AAAI Conference on Weblogs and social Media Washington DC, 2010

[3] Sitaram Asur, Bernardo A. Huberman, Predicting the future with social media, Computing Research Association for the CIFellows Project, 2010, 3

[4] Daniel Gayo-Avello, A balanced survey on election prediction using Twitter data, arXiv, 2012, 4: 1~13

[5] Magnus Lie Hetland, Foundamentals of Python, People's Posts and Telecommunications Pres, 2010, 1: 1~460

[6] Mark Summerfield, Programming in Python 3, People's Posts and Telecommunications Press, 2011, 1: 2~462

[7] Anany Levitin, Fundamentals of Algorithm Analysis and Design, Tsinghua University Press, 2003, 1: 6~430