Crawling and Analysis of Data Based on Social Networking on Stock Comments

Crawling and Analysis of Data Based on Social Networking on Stock Comments

Gang Gao, YuPeng Liu and Gang Bai

Department of Software Engineering, Harbin University of Science and Technology
Email: 962181995@qq.com, flyeagle99@126.com,1978489995@qq.com

Abstract. In this paper, stock data on social networking platforms for the research object, a reptile related technologies .The main purpose of this paper is to design web crawler programs for web sites to crawl stock information and comments. Analysis of crawling information will be used to help investor to do their own investment strategies. The details and application of Python crawler are described in detail. And we have compared with the baseline system to exhibit our crawler's advantages. The work as follows:1.Analysis of the existence of the current social network information during the crawling problem, which leads to the need to achieve the design goal of reptiles; 2.In order to crawl enormous data from different social networks, we constructed several different social networks information suitable for crawling web crawler by using different methods.; 3.Using regular expression to quickly parse large amounts of text to find specific character patterns; extract, edit, replace, or delete text substrings. 4. Climb take the information including stock opening price, stock closing price ,turnover and content comment ; 5.Putting the information into the document for data storage .

1. Introduction

The explosive growth of data available on the World Wide Web has made this the largest publicly accessible data-bank in the world. Therefore, it presents an unprecedented opportunity for data mining and knowledge discovery. Web mining, which is a field of science that aims to discover useful knowledge from information that is available over the internet, can be classified into three categories, depending on the mining goals and the information garnered: web structure mining, web usage mining, and web content mining[1]. As a typical application in the era of Web2.0, social networking services are rapidly becoming popular around the world[2]. The emergence of social networks has pushed social forms and user behaviors on the Internet to the real world, and virtual society and the real world have begun to cross. With the continuous development of social networks and the rapid increase in registered users, more and more researchers are involved in the research of multi-faceted content. According to the data in the 35th Statistical Report on Internet Development in China[3], as of December 2014, the number of Internet users in China reached 649 million, and the number of mobile Internet users reached 557 million. According to relevant reports, Sina Weibo's registered users have exceeded RMB 600 million. In 2015, Sina Weibo's first quarter financial report[4]showed that the number of active users in the first quarter reached 198 million, and the average daily active users were also It reached 89 million.

Foreign researchers mostly conduct data collection for Facebook[5],Twitter[6][7], etc. Researchers set up a network model on the Twitter platform to study the characteristics of the network. Many of the underlying interfaces and libraries of Twitter are public. Use this feature to make statistics on Twitter data. Facebook also gradually opened the API interface. Researchers use the successive interfaces to collect and research Facebook user data. Sina Weibo has become a powerful data-

providing platform for domestic researchers to conduct data mining research and analysis [8].Tan Lian [9]and others proposed an acquisition method that combines web page crawling based on microblogging APIs and web crawlers. However, this method is inferior to web crawlers in acquiring data integrity, and the rate at which API obtains large amounts of user IDs is lower than page acquisition. Huang Yanwei[10]and others studied WeiBo APIs and information acquisition technology based on network data flow. In this technology, simply using the API interface for data collection will result in restricted access, and the amount of collected user WeiBo data is minor.

The remainder of this paper is structured as follows. Section 2 presents the key technical requirements that guided the crawler development. Section 3 presents the strategies and goals of reptile. Section 4 covers the main Crawler features and their implementation details. Section 5 our experimental results are described, as well as a discussion of the performance measures. Finally, Section 6 presents our conclusions and directions for future research.

2. Goals and strategies

In this chapter, we describe the strategies and goals that have guided our crawler implementation. For making our information more accurate and more convincing, we adapt design two different reptiles, which can crawl News Finance Websites and networks respectively. There are three stages:

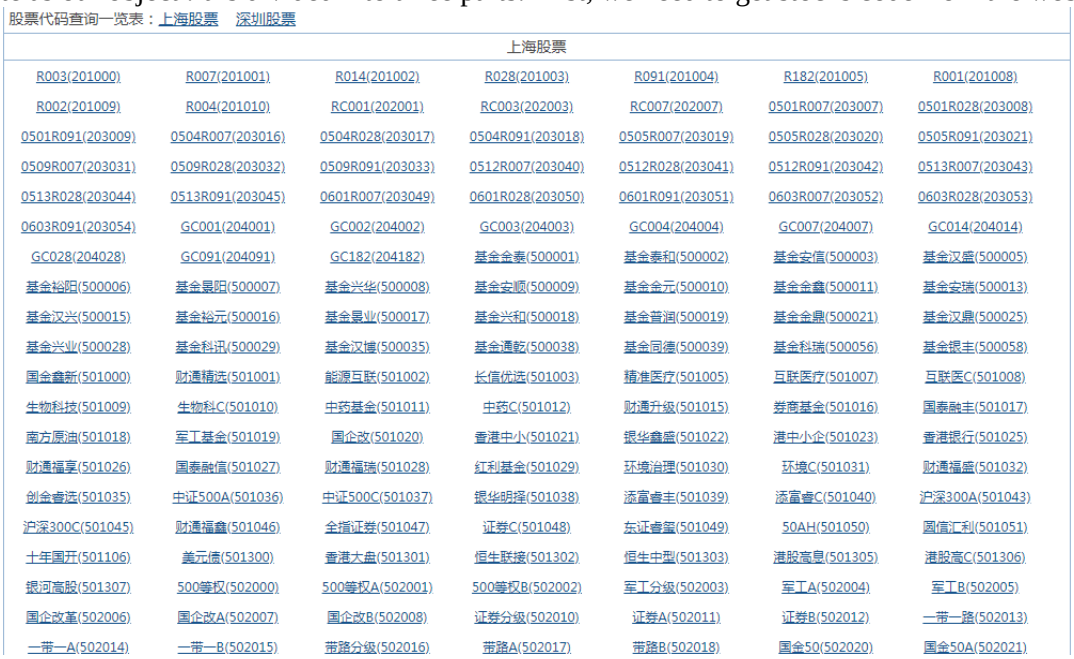
1. First, we need to know about accurate stock information by choosing to crawl these News Finance Website such as NetEase, and hence we can know some comments whether right or wrong and give people a right orientation to follow .In this part ,we mainly use regular expression to get information from website ,we will introduce these details in section 4.

2. Second, we need to crawl comment about stock from Sina Weibo, the biggest network in China like FaceBook and Twitter in America. According to Weibo which has an anti-spider mechanism, so we can not use regular expression simply like crawl News Finance Website, We use Selenium WebDriver to simulate users to scan browser, it can avoid the anti-spider mechanism if we do not crawl the website too fast . And we also will introduce these details in section 4.

3. Finally, we will save these information that crawled from website .And give a conclusion about our experiment.

3. Functional properties and implementation

The first reptile we called reptileNo.1, which crawled News Finance Website, we chosen the NetEase website as our object .It is divided into three parts. First, we need to get stocks code from the website.


股票代码查询一览表：上海股票 深圳股票

上海股票

R003(201000)	R007(201001)	R014(201002)	R028(201003)	R091(201004)	R182(201005)	R001(201008)
R002(201009)	R004(201010)	RC001(202001)	RC003(202003)	RC007(202007)	0501R007(203007)	0501R028(203008)
0501R091(203009)	0504R007(203016)	0504R028(203017)	0504R091(203018)	0505R007(203019)	0505R028(203020)	0505R091(203021)
0509R007(203031)	0509R028(203032)	0509R091(203033)	0512R007(203040)	0512R028(203041)	0512R091(203042)	0513R007(203043)
0513R028(203044)	0513R091(203045)	0601R007(203049)	0601R028(203050)	0601R091(203051)	0603R007(203052)	0603R028(203053)
0603R091(203054)	GC001(204001)	GC002(204002)	GC003(204003)	GC004(204004)	GC007(204007)	GC014(204014)
GC028(204028)	GC091(204091)	GC182(204182)	基金金泰(500001)	基金泰和(500002)	基金安信(500003)	基金汉盛(500005)
基金裕阳(500006)	基金景阳(500007)	基金兴华(500008)	基金安顺(500009)	基金金元(500010)	基金金鑫(500011)	基金安瑞(500013)
基金汉兴(500015)	基金裕元(500016)	基金景业(500017)	基金兴和(500018)	基金普润(500019)	基金金鼎(500021)	基金汉鼎(500025)
基金兴业(500028)	基金科讯(500029)	基金汉博(500035)	基金通乾(500038)	基金同德(500039)	基金科瑞(500056)	基金银丰(500058)
国金鑫新(501000)	财通精选(501001)	能源互联(501002)	长信优选(501003)	精准医疗(501005)	互联医疗(501007)	互联医C(501008)
生物科技(501009)	生物科C(501010)	中药基金(501011)	中药C(501012)	财通升级(501015)	券商基金(501016)	国泰融丰(501017)
南方原油(501018)	军工基金(501019)	国企改(501020)	香港中小(501021)	银华鑫盛(501022)	港中小企(501023)	香港银行(501025)
财通福享(501026)	国泰融信(501027)	财通福瑞(501028)	红利基金(501029)	环境治理(501030)	环境C(501031)	财通福盛(501032)
创金睿选(501035)	中证500A(501036)	中证500C(501037)	银华明辉(501038)	添富睿丰(501039)	添富睿C(501040)	沪深300A(501043)
沪深300C(501045)	财通福鑫(501046)	全指证券(501047)	证券C(501048)	东证睿盛(501049)	50AH(501050)	圆信汇利(501051)
十年国开(501106)	美元债(501300)	香港大盘(501301)	恒生联接(501302)	恒生中型(501303)	港股高息(501305)	港股高C(501306)
银河高股(501307)	500等权(502000)	500等权A(502001)	500等权B(502002)	军工分级(502003)	军工A(502004)	军工B(502005)
国企改革(502006)	国企改A(502007)	国企改B(502008)	证券分级(502010)	证券A(502011)	证券B(502012)	一带一路(502013)
一带一A(502014)	一带一B(502015)	带路分级(502016)	带路A(502017)	带路B(502018)	国金50(502020)	国金50A(502021)

Figure 1. Stocks from website

Through checking from web source codes, we can see stock codes, use the regular expression to get stock code and save in the Python list. Then, Select from the list of stock codes that have been saved, connect NetEase's stock historical data url with stock code to get a new url, use the test tool of Google Chrome, and switch to the network, we can get the request interface when we click to download: http://quotes.money.163.com/trade/lsjysj_zhishu_399301.html?year=2017&season=3

We can change the stock codes from the url to get different historical stock trade information. Requesturl:http://quotes.money.163.com/service/chddata.html?code=1399301&start=20130110&end=20180612&fields=TCLOSE;HIGH;LOW;TOPEN;LCLOSE;CHG;PCHG;VOTURNOVER;VATURNOVER

We can see the code ,start time and the end time clearly from this.

日期	开盘价	最高价	最低价	收盘价	涨跌额	涨跌幅(%)	成交量(股)	成交金额(元)
20170929	156.85	156.86	156.84	156.86	0.03	0.02	832,400.00	81,645,806.76
20170928	156.77	156.84	156.77	156.83	0.08	0.05	2,227,303.00	222,256,647.56
20170927	156.74	156.76	156.73	156.75	0.03	0.02	1,367,506.00	133,178,138.71
20170926	156.73	156.73	156.71	156.72	0.01	0.01	2,563,790.00	249,471,772.07
20170925	156.69	156.73	156.67	156.71	0.08	0.05	3,030,963.00	294,720,276.58
20170922	156.64	156.65	156.46	156.63	0.02	0.01	2,697,666.00	266,397,425.01
20170921	156.59	156.61	156.59	156.61	0.04	0.03	2,860,478.00	274,928,063.58

Figure 2. The trade price of stocks.

Figure 3. The website URL

So, we can get the stock data from finance News websites. Second, we need to crawl network .In this paper, we select the Sina Weibo as our object. Then, we will introduce every steps in next part .And we called this reptile as NO.2.

First, we choose The Shanghai Composite Index as an example in our experiment. So, we aim to crawl comment data from Sina Weibo about this subject . Then, as a crawler, we all know how to check the source code about a website, click the network and then slide the mouse wheel, get the header data and URL by finding the JSION document .

url:https://m.weibo.cn/api/container/getIndex?type=all&queryVal=%E4%B8%8A%E8%AF%81%E6%8C%87%E6%95%B0&featurecode=20000320&luicode=10000011&lfid=106003type%3D1&title=%E4%B8%8A%E8%AF%81%E6%8C%87%E6%95%B0&containerid=100103type%3D1%26q%3D%E4%B8%8A%E8%AF%81%E6%8C%87%E6%95%B0&page=3

By changing the page in this, we can get different pages .And then, we open the URL, and we can see the figure 4. Using “response. json()[‘data’][‘cards’][0][‘card_group’][]” to get we want. The comments are saved in text label. Finally, we have saved these in txt documents, and prepared for data cleaning.

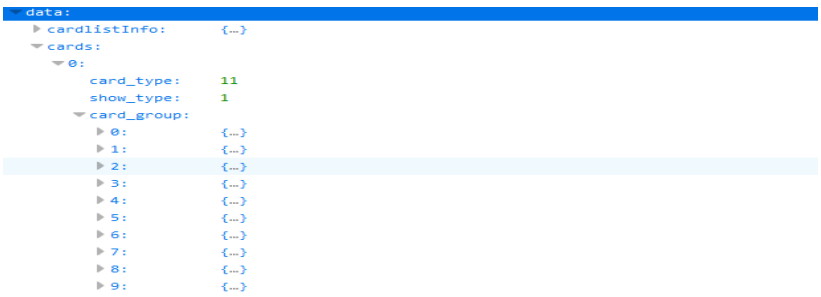


Figure 4. The website URL

IOP Conf. Series: Earth and Environmental Science **234** (2019) 012093 doi:10.1088/1755-1315/234/1/012093

```
url='https://m.weibo.cn/api/container/getIndex?type=all&queryVal=%E4%B8%8A%E8%AF%81%E6%8C%87%E6%95%B0&featurecode=20000320&luicode=10000011&lfi
ii=0
while HTML(ii).status_code==200:
    ii=ii+1
   try:
        for i in range(0,len(HTML(ii).json()['data']['cards'][0]['card_group'])):
            data1=[(r.json()['data']['cards'][0]['card_group'][i]['mblog']['created_at']  ,
                  r.json()['data']['cards'][0]['card_group'][i]['mblog']['id']  ,
                  r.json()['data']['cards'][0]['card_group'][i]['mblog']['text'],
                  r.json()['data']['cards'][0]['card_group'][i]['mblog']['source'])]
            print(data1)
    catch:
        continue
```

Figure 5. The kernel code in the reptile

4. Experimental results and discussion

4.1 Running the experiment

We run our experiments on a common PC, with 6 GB RAM and a 100 GB SSD hard drive, where the CPU is a core i7 and the operating system is Microsoft Windows 10. Pages are collected from the NetEase finance News website and Sina Weibo website.

Figure 6. The comments from Sina Weibo.But there are many pictures and websites' links, so we may will do data cleaning for these documents.

Figure 7. The stocks price from NetEase news website.

During crawling these data from different websites, we also found some difficult in crawling .First, we found it is difficult for us to crawl Weibo website, because there is a significant anti-reptile in Weibo website. So, we have chosen Weibo platform on phone, it is same as website platform, and it is prone to crawl for us . Then, we tried to crawl comments under one stock, and we succeeded although there are many useless data and unnecessary links in our documents. The resulting statistics for the experiment are presented in figure 6.

The stock crawler mainly crawls some stock features. All stock codes are written in the html page, and the stock code is extracted by using regular expressions. The search speed is faster and accurate, and all the codes are obtained accurately. The cycle time complexity is O(n). For stock information crawling, using string splicing can get all the stock information, it is relatively easy to get stock information on Netease Finance. Its opposition to web crawlers is not very large, we can use time delays to ensure our method is friendly for the server. Sina Weibo stock comment, due to Sina Weibo comment area data is dynamic, can not be obtained by conventional methods. We need to manually grab the packet and get the URL, then search through it, maybe sometimes there will be some empty pages in it, which is wasteful of time and slows down the efficiency of crawling . We can extract information by using json data bank, whereby improves the time efficiency, because the comment data area is presented in json format.

4.2 Evaluating Crawler speed performance

To make a fair comparison, we have selected two similar data platforms for our crawlers. Figure 8 shows the Reptile No.1 crawled the amount of data during a given time period. Similarly, figure 9 presents the Reptile No.2 crawled the amount of data during a given time period.

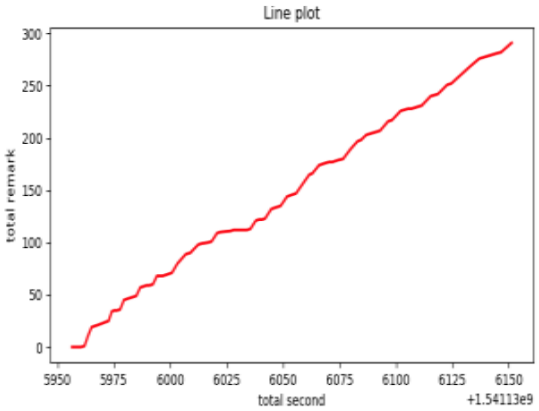


Figure 8. The performance of Reptile No.1

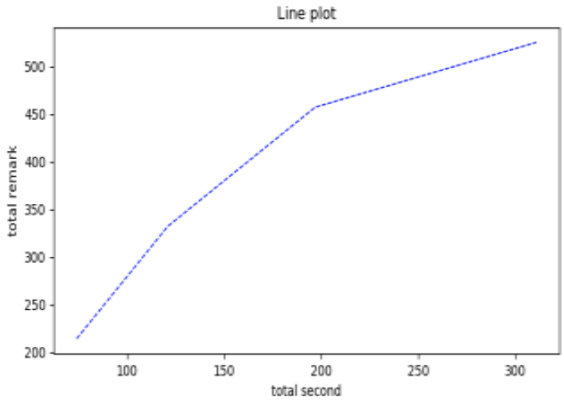


Figure 9. The performance of Reptile No.2

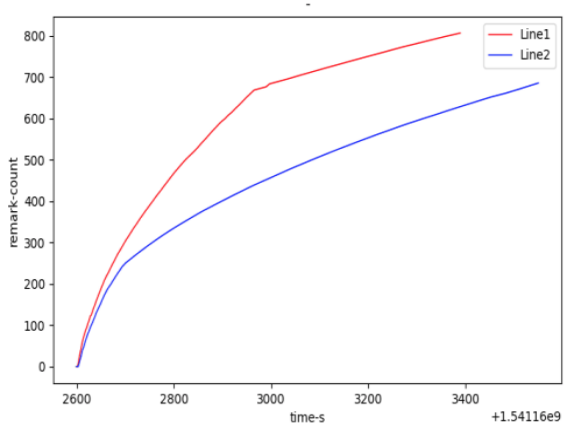


Figure 10. Reptile No.2 achieves a high performance compared to baseline system during the given time. (Line 1 is Reptile No.1and Line 2 is baseline system)

It is easy to see that the ability of our crawlers can crawl how much data during the given time. In order to know the advantages of our reptiles compared to the current reptiles, we have compared with the baseline system, which is a stable reptile on the internet, and did a comparing image for them. Figure 10 depicts this.

It is obvious that our crawler, Reptile No.2, is prior to the baseline system. In fact, multi-threading enables efficient use of available resources. So we have taken the multi-threading to improve the speed. Although the crawler speed is considered as a performance factor, a crawler should also be polite to web servers, such that it does not overload the server with frequent requests within a short space of time. Therefore, there should be sufficient delays between consecutive requests to the same server.

5. Conclusion and future work

With the rapid development of various social networks in recent years, especially Weibo has developed into an important social media. Getting news, current affairs, self-expression, social sharing and other content from Weibo has become a part of daily life. In this paper, we presented our reptiles which have crawled different data not only from Sina Weibo, also from NetEase finance News website. For different data requirements, designed different reptiles as our papers. We intend to improve its speed and effect. We also plan to extend its features and functionalities by implementing focused crawling and automatic content detection and extraction.

6. References

[1] Markov Z, Larose DT. Data Mining the Web: Uncovering Patterns in Web Content, Structure, and Usage. John Wiley & Sons; 2007.
[2] WANG Yuan-zhuo,JIN Xiao-long,CHENG Xue-qi.Network Big Data:Present and Future[J].Chinese Journal of Computers, 2013,36(6):1125-1138.(in Chinese)
[3] Wang Yuan, Zhuo Zhu, Xiaolong Xiao, Cheng Xueqi. Network Big Data: Present Status and Prospects [J]. Computer Science, 2013, 36(6): 1125-1138.
[4] China Internet Information Center. The 35th Report on China's Internet Development [EB/OL].[2015-02-03]. http://www.cnnic.net.cn/hlwfzyj/Hlwxzbg/hlwtjbg/201502/t20150203_51634.htm.
[5] Sina Technology. Financial Report for the First Quarter of 2015 [EB/OL].[2015-05-15]. http://tech.sina.com.cn/i/2015-05-15/doc-iavxeafs7518570 .Shtml.
[6] STAM K M, CAMERON G T, STAM A, et al.Stam Sociometric Attractiveness on Facebook[J].Proceedings of the International Conference on Information Manag, 2014, 6(6): 180-188.
[7] ALLOWAY T P, ALLOWAY R G. The impact of engagement with social networking sites (SNSs) on cognitive skills [J].Computers in Human Behavior, 2012, 28(5): 1748-1754.
[8] DING Zhao-Yun, JIA Yan, ZHOU Bin.Survey of Data Minging for Microblogs [J] .Journal of Computer Research and Development, 2014,51 (4):. 691-706 (in Chinese)
[9] Ding Zhaoyun, Jia Yan, Zhou Bin. Microblogging data mining research and research [J]. Computer Research and Development, 2014, 51(4): 691-706.
[10] LI D, NIU J, QIU M, et al.Sentiment analysis on Weibo data [C]// 2014IEEE Computing, Communications and IT Applications Conference (ComComAp). IEEE, 2014: 249-254.
[11] LIAN Jie, ZHOU Xin, LIU Yun.SINA microblog data retrieval [J].J T sing hua Univ(Sci & Tech),2011,51(10):1300-1305.(in Chinese)
[12] Lian Jie, Zhou Xin, Liu Yun . New Wave Microblogging Data Digging Method [J]. Qing Hua Da Xue Bao (Natural Science Edition), 2011, 51(10):1300-1305.
[13] HUANG Yan-wei, LIU Jia-yong studied on Sina WeiBo Data Acquisition Technology[J].Information Security and Communication Security,2013(6):71-73.(in Chinese)
[14] Huang Yan-wei, Liu Jia-yong. Sina WeiBo Data Acquisition Technical Research [J]. Information Security and Communication Security, 2013(6):71-73.