Anti-Crawler Strategy and Distributed Crawler Based on Hadoop

HongRu Wang
School of Computer Science,
Communication Unversity of China,
Beijing, P.R. China
rule2018@163.com

ChunFang Li
School of Computer Science,
Communication Unversity of China,
Beijing, P.R. China
LCF@cuc.edu.cn

LingFei Zhang
School of Computer Science,
Communication Unversity of China,
Beijing, P.R. China
echozhang1631@163.com

MinYong Shi
School of Computer Science,
Communication Unversity of China,
Beijing, P.R. China
myshi@cuc.edu.cn

***Abstract*—With the exponential growth of network information resource, distributed Web Crawler was introduced for fetching massive web pages. However, when many websites designed a series of Anti-crawlers tactics to disturb Web Crawlers, it is our first imperative to deal with these strategies. We first elaborated on Web-Crawlers and Anti-Crawlers strategies.At the same time, this paper mainly proposed some effective measures to guarantee efficient web-crawling when faced with Anti-crawler tactics.Then we designed a experiment to analyze possible influencing factors.Finally, the future direction about distributed Web Crawels which runs on Hadoop platform was proposed by this paper.**

Keywords-Anti-crawlers; Distributed Crawler; Nutch; Hadoop

I. INTRODUCTION

Along with the rapid development of the Internet , web crawlers , as applications or services to find useful and related information on the web, are becoming increasingly important as the main means of locating information. However, recent surveys have shown that many websites gradually take some Anti-Web Crawler strategies in their inner system due to this trend .According to these strict measures,it increased difficulties of crawling information. In this case, several ways to deal with Anti-Web Crawler strategies come into being.

As the ability of traditional simple stand-alone web crawlers has been unable to keep pace with the growth of information on, distributed crawlers become a new boom. So far, there are many academic research institutions and enterprises in the study of related issues. Alibaba Group Holdings Corporation applied for a patent about "User attribute value calculation method and computing device based on user browsing behavior"[1]. In 2017, Lei Zhang and JianJun Han published a paper related to A distributed crawler system and periodic incremental crawl method is currently in use in business [2].

In this paper, we mainly focus on the principles of crawlers and a series of methods to deal with anti-web crawlers strategies. On this basis, we design and implement an experiment to test several different ways to deal with anti-Crawlers.Finally, this paper introduces some knowledges about distributed web crawlers based on the frame of the Hadoop Distributes and cloud computing to show the bright future of web crawlers.

II. WEB CRAWLER AND ANTI-CRAWLER

A. Web Crawler

A Web crawler, sometimes called a spider, is an Internet bot that systematically browses the World Wide Web typically for the purpose of Web indexing (web spidering).Sometimes it is applied for updating web content by web search engines in some sites[3]. Web crawlers are an essential component of all search engines, and are increasingly becoming important in data mining and other indexing applications.

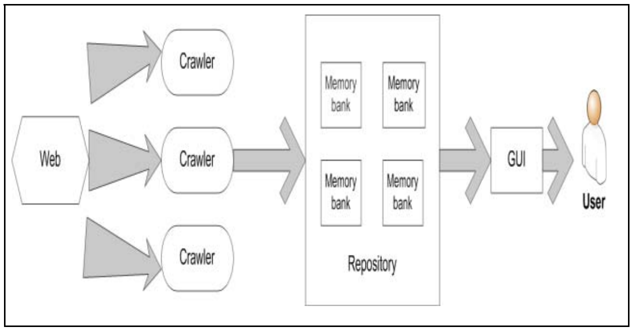

Figure 1. Crawler physical structure

978-1-5386-4794-3/18/$31.00 ©2018 IEEE

Simple single-threaded web crawler are mainly composed of these modules: DNS, web crawling, web parsing, web processing and a series of URL-related modules.

- DNS: This module is a URL addressing module and its function is getting the web page based on URL.
- Web Crawling: Getting a specified resource on the Internet requires crawling the webpage, and the captured network resources are stored on the user's computer as a data stream.
- Web Parsing: The problem to be solved by this module is to extract structured information and denoising the webpage for a specific webpage.
- Web Processing: This module parses the webpage information processed by the web parsing module again to remove duplicate documents, which saves space, users' time and improves the search quality.
- URL-Related Modules: It includes URL extraction module, URL library and URL library to be crawled.

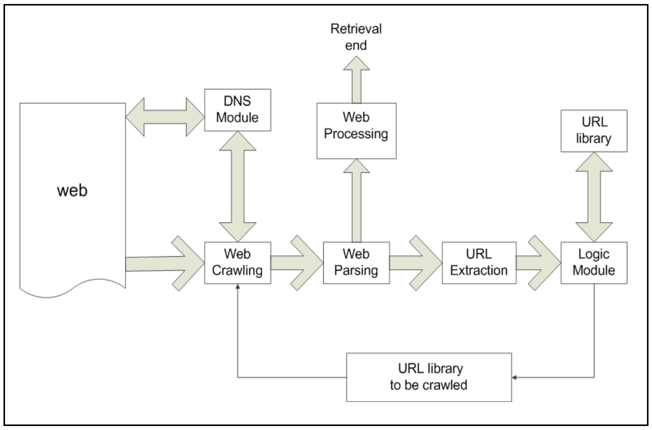

Figure 2. The principle of Web Crawler

B. Anti-Crawler

When it comes to using web spider to crawl data , we have to think about the issues which limit speed to crawl and some about how to address anti-crawlers strategies used by these websites.Lots of websites use different ways to protect their data and all we need to do is avoid these strategies and acquire the data we need.

First of all, if we want to crawl data,a computer is essential.our pc only have one ip address,so if we have been using the same computer to crawler data,some websites will recognize the ip and ban it for a period time[4].More serious thing is that the website will prohibit the ip permanently.

We can bulid our own ip proxy pool or find some public ips.The second method always be useless cos most these ips are unvalid.So in this process,we must check whether the ip adress is valid which is exported from proxy pool.

Then some websites ususlly build their anti-crawler system base on cookies,this means it will remember our cookie when we visit the website.So the reponsive way is maintain a cookie pool and upgrade regularly,and take cookie randomly form it when you crawl data.Another means used by the website to against the crawler is the verification code.

We can use the open source Tesseract-OCR system or other ways to download and identify the picture about the verification code, and pass the characters to spider sysytem for simulated landing.Of course we also can upload the code to the coding platform for identification and update the verification code again until successful.

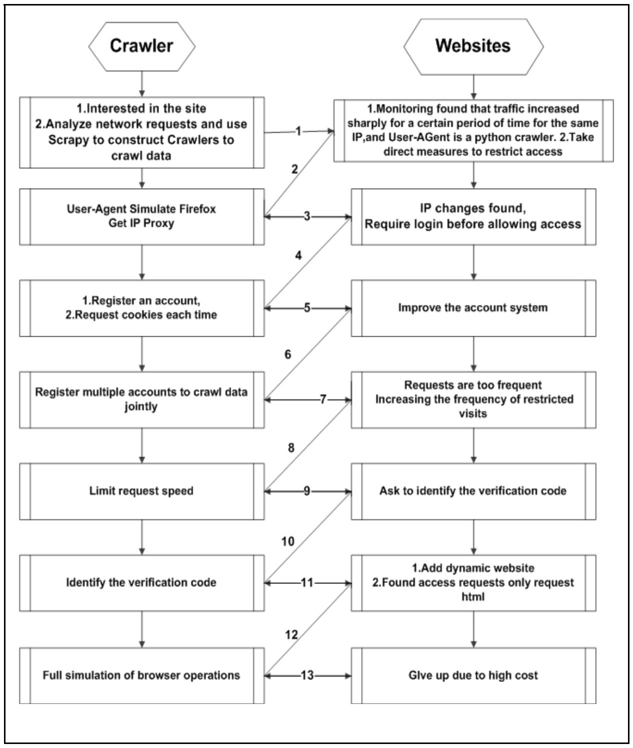
Crawler
Websites
1.Interested in the site
2.Analyze network requests and use Scrapy to construct Crawlers to crawl data
1.Monitoring found that traffic increased sharply for a certain period of time for the same IP,and User-AGent is a python crawler. 2.Take direct measures to restrict access
User-Agent Simulate Firefox
Get IP Proxy
IP changes found,
Require login before allowing access
1.Register an account,
2.Request cookies each time
Improve the account system
Register multiple accounts to crawl data jointly
Requests are too frequent
Increasing the frequency of restricted visits
Limit request speed
Ask to identify the verification code
Identify the verification code
1.Add dynamic website
2.Found access requests only request html
Full simulation of browser operations
GIve up due to high cost
1
2
3
4
5
6
7
8
9
10
11
12
13

Figure 3. Anti-crawler strategies

Sometimes websites will use different kinds of ways to prohibit us, but there are lots of other ways for us to use for anti-anti-spider,including disguised as a browser、multi-threaded access、multi-process crawl and limit the speed of visit,etc.In some cases,we need add both User-agent and Referer into header to pretend to be browser.In the next experiment, we used different methods to observe and stop the site's anti-crawlers strategy.

III. EXPERIMENT

A. Experiment Overview

According to the introduction above, our research mainly focus on how to deal with anti-crawler strategies and crawl information effectively .Nowadays,the main anti-crawler strategies can be divided into two categories --anti-crawler by Headers requested by user and anti-crawler based on behavior of users.In this section,firstly we propose three strategies and apply them to a specific distributed crawler experiment.

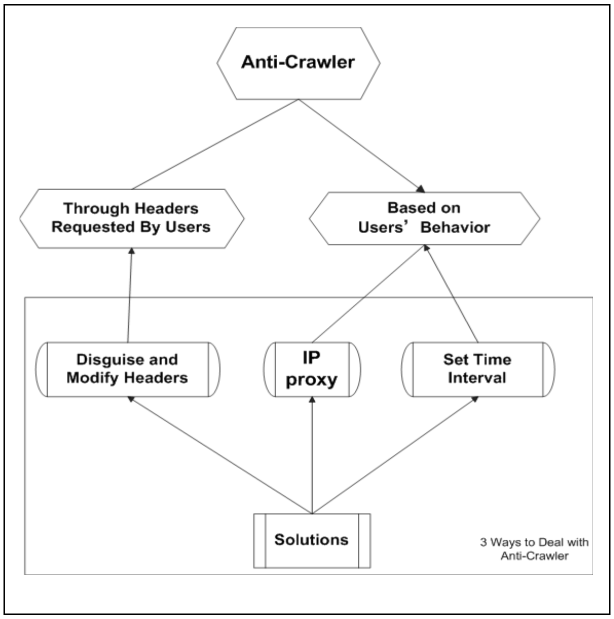
Anti-Crawler
Through Headers Requested By Users
Based on Users' Behavior
Disguise and Modify Headers
IP proxy
Set Time Interval
Solutions
3 Ways to Deal with Anti-Crawler

Figure 4. Solutions to deal with anti-crawler

The method adopted by the experiment to cope with the corresponding anti-crawler strategy is shown in Figure 4. With the rapid growth of the amount of web data, it is a great of necessity for us to build a robust system which can handle more data and can be used in more situations.Except distributed crawler, the other ways we could apply are multithreading and proxy ip,etc.First of all, we crawl some ip adress from third-party websites and check its validity by using it to visit websites such as www.baidu.com and so on, after that, we save those effective ip adress as a list to use at next step. And then we take one ip out from the list randomly and termly while we are crawling data by using proxyhandler which is provided by urbllib2-----a package used in python.In order to improve efficient and energy,we applied multithreading to get ip address and data we need especilly when we need to visit different websites.For instance,the ip address are saved in eight websites and the constructure of html of those websites is totally same.so we can use same regular expression and establish eight threads to get ip from eight websites.

Another method applied in our experiment is disguise and modify headers.Some websites regularly detect third-party foreign chains to prevent crawlers from crawling large amounts of data,pretend crawler as normal browser access is a way that we can use to solve it by using build_opener embeded in urllib2 module and adding headers in this opener.

B. *Experiment Process*

What's more,in order to verify the effectiveness of these method,we change some parameters for analysis and comparison.We first use multi-ip address on a single machine to crawl data, we select different parameters to estimate the threshold of the target website by changing the number of threads and the total number of samples that need to be crawled each thread(each sample contains 10 different dimensions of data whose datatypes include texts, numbers, etc.) in this process, and we also adjust the waiting time for converting IP addresses between different threads according to the actual situation. Finally,as the table is given, by repeatedly testing the website ceiling, we found that the most important factors are samples for each thread and sleep time during a break.

Table 1. Influential factors

Threads	Samples for each thread	Sleep Times(s)	End
10	70	20	10
10	140	20	2
10	140	60	4
15	70	60	6
20	60	60	12

When we crawl datas at a great speed,the website may take compulsory measures to stop our crawling.So it's quite feasible for us to end crawling and adjust above factors to restart it.We use the special value,the number of thread when we need to end and restart crawling,as an indicator to evaluate our Crawler.

As is shown in the below figure, we have the following procedures:

- Keep the number of threads and sleep time unchanged.Adjust the total number of samples in each thread.As the number of samples increases, the number of threads when terminated is less. In other words, the more samples that need to be crawled in each thread, the easier it is for the website to adopt Anti-Crawler strategies.
- Change sleep times and keep other variables unchanged.It's found that the shorter the sleep time, the easier it was found and prevented by sites.
- Only make changes to the number of thread.After the test,we speculate that the more threads that are assigned to the crawler system, the easier it is to envade the web's Anti-Crawler methods.

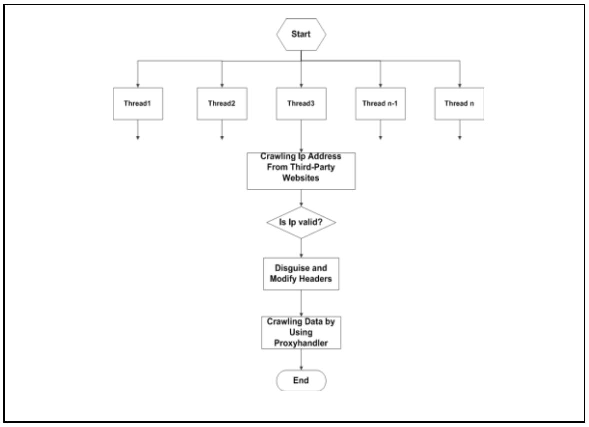
Start
Thread1
Thread2
Thread3
Thread n-1
Thread n
Crawling Ip Address From Third-Party Websites
Is Ip valid?
Disguise and Modify Headers
Crawling Data by Using Proxyhandler
End

Figure 5. Experimental flow chart

IV. DISTRIBUTED CRAWLER ON HADOOP

A. Basic Information

Consider all kinds of needs of all users, there will be a variety of bottlenecks in some cases for a stand-alone web crawler. In the situation that cloud computing has become one of the important trends in the future development, there are many cloud platforms to help us to better solve this problem such as AWS、Hadoop.Next we will introduce the widely applied distributed platform –Hadoop[5].

B. Hadoop--- a Distributed Platform

Hadoop is a sub-project of Lucene, which was originally part of the "Nutch" project and was detached from Nutch in early 2006.Then it became an independent project. Hadoop is not just a distributed file system for storage alone, but a framework designed to execute distributed applications on large clusters of commodity hardware devices.

There are two parts included in Hadoop distributed platform[6]:Hadoop distributed file system--HDFS and distributed calculation framework--Map/Reduce.As is shown in Figure 6.

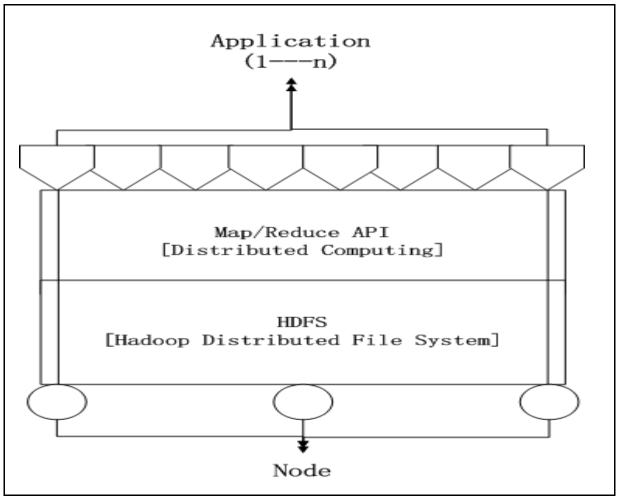

Figure 6. Hadoop Distributed Platform

As we know,distributed file system is the underlying support for large-scale distributed computing.Administrators can organize shared folders on different servers together into a single directory tree. It appears to the user that all shared files are stored in just one place and that they can access files or folders distributed over the network by simply accessing a shared DFS root without having to know the actual physical location of the files.

- HDFS, like other distributed file systems, has several basic features. Firstly,it has a single namespace for the entire cluster. The second is data consistency. What's more, the file will be divided into multiple file blocks, each block is allocated to store on the data node, and will be copied by the file block to ensure data security according to the configuration.
- Map/Reduce, as a simplified distributed programming model , allows programs to be distributed and executed in a cluster of common machines in parallel.

C. Distributed Crawler Based on Hadoop

Distributed Web crawler is developed from the traditional reptile, so it works and the basic structure is similar to the traditional Web crawler. It can be used as a combination of multiple Web crawler macros. One of the most useful and popular distributed web crawler is Nutch which is an open source Java implementation of the search engine. It provides all the tools we need to run our own search engine, including full-text search and web crawlers[7].

First deployed Hadoop and Hbase, and then configure nutch, we must pay attention to the match between versions. If you do not want to store data in Hbase, you can also choose Mongodb, Mysql and so on.

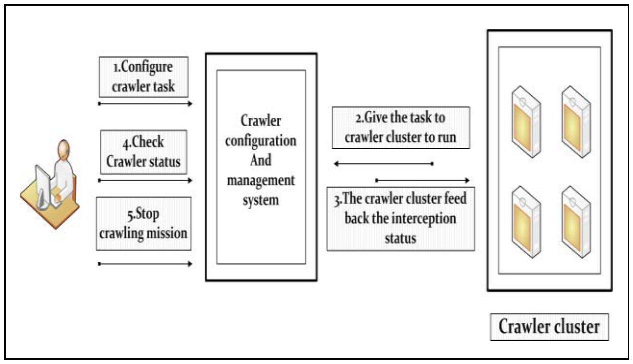
1.Configure crawler task
4.Check Crawler status
5.Stop crawling mission
Crawler configuration And management system
2.Give the task to crawler cluster to run
3.The crawler cluster feed back the interception status
Crawler cluster

Figure 7. Crawler Clusters

Based on Nutch and Hadoop, the web crawler architecture is a typical distributed off-line batch processing architecture with very good throughput and crawl performance and offers a large number of configuration customization options.Distributed database architecture base on Hbase and Hadoop, is a typical distributed online real-time random read and write architecture. Strong horizontal scalability, support for billions of rows and millions of columns, to be able to write data submitted by web crawlers in real time.

We divide this distributed reptile into four modules: information acquisition module, analysis module, storage module, user interface module. First of all, the information acquisiyion module crawls the html file from the Internet and saves the downloaded file in Hadoop's HDFS, and then passes some screenings according to demand scene to eliminate some redundant content. Then in the analysis module, the preliminary screening content Processing, and the data according to different attributes or according to different needs to be stored in Hbase, the user can browse or search the content that saved in the database on GUI. In this process, the information acquisition, storage modules are based on Hadoop's distributed computing framework implemented by MapReduce[8].

V. SUMMARY

We have proposed some strategies based on distributed thoughts which aims to handle Anti-Crawler tactics.The experiment results shows that some plausible factors influence the Web-Crawels performance .Distributed Web-Crawlers which runs on Hadoop platforms is an important component of contemporary Web-Crawler.In order to make a deep study of distributed Web-Crawler,We will reaserch on knowledges about Hadoop platforms and cloud computing applied in Web Crawler after more experiments.

ACKNOWLEDGMENT

This paper is partly supported by “Key Cultivation Engineering Project of Communication University of China (Project number: 3132017XNG1606 and 3132017XNG1719)”, “the Excellent Young Teachers Training Project (the second level, Project number: YXJS201508)”, “Cultural technological innovation project of Ministry of Culture of P. R. China (Project number: 2014-12)”. The research work was also supported by “Chaoyang District Science and Technology Project (CYXC1504)”,“The Scientific Research Training Program of Beijing College Students”(Project number:GJ170803).

REFERENCES

[1] Hui Li,JunXing Gao,Dong Sheng “User attribute value calculation method and computing device based on user browsing behavior” [P]. China Patent: CN107122367A, 2017-09-01.

[2] Lei Zhang,JianJun Han “A distributed crawler system and periodic incremental crawl method” [P]. China Patent: CN107193960A, 2017-09-22.

[3] Cheslav Zhdanovich, Michail Mamonov,Maciej Kuboń,Jan Radosław Kamiński. Effect of Steering Gear Parameters of Crawler Tractor Cornering Ability[J]. Agricultural Engineering,2016,20(2):.

[4] JiXiong Yang “Web data anti-crawling methods and systems” [P]. China Patent: CN107220291A, 2017-09-29.

[5] JianKun Yu. “A Distributed Web Crawler Model based on Cloud Computing[A].”(ITOEC 2016)[C].2016:4.

[6] Linping Su. “Web Crawler Model of Fetching Data Speedily Based on Hadoop Distributed System[A].” The Institute of Electrical and Electronics Engineers、IEEE Beijing Section.Proceedings of 2016 IEEE 7th International Conference on Software Engineering and Service Science（ICSESS 2016）[C].The Institute of Electrical and Electronics Engineers、IEEE Beijing Section:,2016:5.

[7] JingYu Gao,KeKe Liu “A distributed web crawler system and information crawling method” [P]. China Patent: CN107066569A,2017-08-18

[8] Xiaochen Zhang. “Optimization of Distributed Crawler under Hadoop[A].” (ICETA 2015)[C], 2015:6.