ATLANTIS PRESS

Advances in Computer Science Research, volume 83

8th International Conference on Social Network, Communication and Education (SNCE 2018)

Analysis and Research of Distributed Network Crawler based on Cloud Computing Hadoop Platform

Hongsheng Xu[1,2 a *], Ganglong Fan[1,2] and Ke Li[1,2]

[1]Luoyang Normal University, Luoyang, 471934, China

[2]Henan key Laboratory for Big Data Processing & Analytics of Electronic Commerce, Luoyang, 471934, China

[a]85660190@qq.com

* The Corresponding Author

Keywords: Cloud computing; Network crawler; Hadoop platform; Data processing; Distributed network

Abstract. Cloud computing is a new way of network service, which transforms the traditional desktop task processing into the network based task processing. Hadoop is a Java - based software framework for distributed intensive data processing and data analysis. Web crawler is a program or script that automatically captures web information according to certain rules of Internet access. This paper presents analysis and research of distributed network crawler based on cloud computing Hadoop platform. Distributed network crawling nodes can be divided into four parts: network crawler module, node information maintenance module, task allocation module, node communication module.

Introduction

In the early days of the Internet, the number of sites is relatively small, the amount of data information, search easier. However, with the development of Internet blowout, ordinary Internet users want to find their required data like looking for a needle in the sea of data, in order to meet the needs of professional search site public information retrieval demand came into being.

Web crawler is a program or script for automatically grabbing web resources according to certain rules. It has been widely used in the field of Internet [1]. The search engine uses Web crawler to grab Web pages, documents and even pictures, audio, video and other resources. With the rapid development of the network, the World Wide Web has become a carrier of large amount of information. How to extract and utilize this information effectively becomes a huge challenge.

The first is the search based on link analysis. In 90s, the foreign search engine developers have started to work as a model of social network, the World Wide Web is simulated. Experts through the network relationship between man and society between the designed and developed a hyperlink relationship network between pages. At the same time they were also surprised to find and the highest similarity in the traditional citation. So you can control through the analysis conclusion, starting from the point of the network, will be able to a large number of Internet web page classification. In early 2002, and appeared the most primitive search system based on link.

The second is the search based on content analysis. Compared with the analysis of link based search methods, which is a breakthrough in the search technology, they adopted a new way of thinking for the establishment of a theme thesaurus. When users search in the professional field, and it is combine the retrieval thesaurus and crawler [2]. Due to the change of search angle, this new technology gradually began to be of concern to the people. In 90s, the Fish Search System as the first search system based on content analysis has been developed.

Search engine is a set of information retrieval system in WWW network environment. It usually has two different working ways: one is classified directory retrieval, which collects the resources in the Internet. According to the different types of resources they provide, they are divided into different directories, and then classified layer by layer, people can find the information they want to enter layer by layer according to their classification, and then they can finally reach their destination. Find the

information you want; The other is keyword based search, which allows users to enter a variety of keywords in a logical combination. According to these keywords, the search engine computer finds the address of the resources required by the user, and then feeds back to the user all the URLs containing the keyword information and the links to these URLs according to certain rules.

The main problem is the development of web crawler and anti blocking performance. In many cases, using high frequency data capture is feasible, the premise is the target site without using any anti climbing measures (access frequency restrict, firewall, authentication code.); more often, valuable information must be accompanied by strict anti climb the measures, once the IP is closed, what components are gone. You have to maintain a proxy IP pool to solve this problem, of course, it also brings the agent IP stability and speed of the problem, these problems are inevitable problems, according to the specific situation we need to adopt corresponding measures, to the maximum the crawler crawling task completion limit [3].

Hadoop is Dougcutting, based on the Java implementation of GFS and MapReduee , and has become part of the famous Lucene project. It was originally part of the Nutch project and was separated from Nutch as an independent project in early 2006. Hadoop is not a simple distributed file system for storage, but a framework designed to execute distributed applications on a large cluster of common hardware devices.

Design and Implementation of Network Crawler

Web crawler is a Web program. According to certain rules, it automatically crawls the Web pages in the World Wide Web, and stores key words in the web pages into keywords database. Users can search pages or related information pages.

Traditional web crawler by climbing URL library from the library URL not to climb, and it is the theme crawler thread module and content extraction module. Its working principle as shown in Figure 1, the web crawler from one or several "initial URl, the initial URL on the page and page information from the web by the theme crawler thread module will get new URL to be stored in the queue URL climb, will get to the web page content extraction module, content extraction module will be visited in URL have been crawling in the URL library, the web page information information stored in the database, to meet certain conditions. Stop by a web page filtering algorithm to filter out the irrelevant URL, follow certain scheduling strategy selection grab next URL [4].

In 1989, during Christmas, the famous Guido van Rossum had nothing to do during Christmas, and to pass the time, he wrote a programming language. This is python. Today, there are hundreds of programming languages around the world, but about 20 of them are the best. The Python language has been around in the last decade. It has always stood in the top 10 of TIOBE, so it's a very good programming language. It's also an evergreen language in the programming language compared to Python. Get closer to hardware, as is shown by equation (1).

$$y(k) = -\sum_{i=2}^{4} den(i)y(c) + \sum_{i=2}^{4} num(i)u(c) \tag{1}$$

Therefore, when programmers needs to write programs that are more demanding in terms of speed and performance. They tend to prefer C, and Python, as a high-level language for writing applications, has a rich and complex base of code that includes files, networks, databases, and text. GUI and so on, so programming through Python is a very simple process, because there is a lot of ready-made code to use. There is no need for programmers to write from scratch [5]. Overall, Python code is simple and elegant. There are many types of applications that are suitable for python development, such as web applications. Script tasks and so on.

The traditional network crawler starts with the URL (Universal Resource Locator uniform resource locator) of one or several initial web pages to obtain the URL on the original webpage. In the

process of grabbing the web page, a new URL is continuously extracted from the current page surface and put into the queue until a certain condition of the system is satisfied.

At the present stage, the network crawler has been developed into an intelligent tool which comprehensively uses a plurality of methods, such as webpage data extraction, machine learning, data mining, semantic understanding and the like. Network crawler security concerns: Because the network crawler' s strategy is as many as possible ' climbing ' sites , as much as possible , access to the page will be made as much as possible according to a specific policy , the network bandwidth is occupied and the processing overhead of the Web server is increased , and the station of many small sites finds that when the Web crawler is patronizing , the access flow will increase significantly. The malicious user can use the crawler program to launch a DoS attack on the Web site, and the resource is exhausted to provide normal service. The malicious user may also grab various sensitive data through the network crawler for improper use.

Today, the focused crawler technology has achieved great development and progress, the typical foreign system including CORA, IBM Focused Crawler.CORA is from the Carnegie Mellon University A.K.McCallum and M.Nigam et al in 1999 on a topic in computer science design search engine.CORA by mechanical cognitive way, the main target is related with the computer the theme of the content, the classification of user needs through the contents of the principle of implicit Malfoy. Although the CORA address and the subject analysis ability is insufficient, also do not have the ability to analyze the web page, it made great achievements in the aspects of the automatic collection of resources but it still cannot be denied.

In order to improve the search engine users search query understanding, there must be a good search query language, in order to overcome the disadvantages of keyword search and directory query, has now appeared in natural language intelligent answering [6]. Users can input simple questions, such as "how can kill virus of computer?" in the search engine. After analyzing the structure and content of questions, or directly gives the answer, then select or guide the user can choose from several problems. The natural language advantage is that a network of communication is more humane, the two is to make the query more convenient, direct and effective.

Excellent representative link evaluation of search engine based on Googel, the original "link evaluation system" is based on such an understanding, the importance of a page depends on the number of other pages link to it, especially some has been identified as "important" link number..

At present, the demand for reptiles is exploding, which is the new normal form of current Internet innovation and big data times [7]. The train and eight - claw fish and other teams have seen this and take the lead in developing relatively complete reptiles. Most users don't know the technology. But more users don't know the technology. Most users don't have this out - out gap. I also believe that the technology gate will continue to gather and form a relatively independent community , and P2P ' s community platform will provide more unobstructed communication channels for crawler developers and reptiles.

Analysis of Distributed Network Crawler based on cloud Computing Hadoop Platform

Cloud computing is a new network service, it will change to the desktop task processing as the core of the traditional task processing in the network as the core, it uses the network to achieve all the tasks you want to accomplish, to make the network become the integrated delivery service, connected computing power and information, realize the on-demand computing, many people work together.

The basic principle is: the use of non local or remote server (cluster) distributed computer, to provide services for Internet users (computing, storage, software and hardware and other services), which makes the user resources can be switched to the needs of the application, according to demand access to computers and storage systems, thereby reducing the cost [8]. The real implementation of the on-demand computing, and it is effectively improving the utilization efficiency of hardware resources.

Big data is different from the traditional type of data; it may be composed of TB or PB information, including structured data, including text, multimedia and other unstructured data. These data types are lack of consistency, can not be effectively stored on large data storage technology makes the standard, and we cannot use the traditional SAN method and server to effectively store and deal with huge amount of data.

These are decided by the "big data" to different processing methods, Hadoop is widely used in large data processing technology. Hadoop is a software framework based on the analysis of the distributed Java intensive data processing and data technology [9]. The heuristic framework elaborated in the 2004 white paper to a great extent by Google MapReduce.

Many Web servers in the Internet will return a list of directories when the client requests a directory that is not the default page in the site. The list of directories typically includes directories and file links available to the user, through which malicious users can access the next layer of directory and files in the current directory. As a result , malicious users can get a lot of useful information , such as the site ' s directory structure , sensitive files , password files , database files , and so on.

Web crawlers are a program or script that automatically captures web information according to certain rules of Internet access. The network retrieval function plays a role in the demand of content retrieval brought by the explosive development of Internet content. The search engine is constantly developing, people's needs are constantly improved, and network information search has become the content of people every day. How to enable the search engine to satisfy people's needs at all times? The original search function is realized by means of index station, so that there are network robots.

Experiments and Analysis

The method comprises the following steps : firstly , establishing a URL task list , namely an initial URL seed , sending a request to a DNS cache by a URL task list , sending an initial URL seed in the URL list to the DNS by the DNS cache , searching the page according to a certain algorithm and an ordering mode by the DNS according to an http protocol , and then sending the obtained URL to the DNS cache according to a certain algorithm and an ordering mode , and then sending the obtained URL to the DNS cache according to a certain algorithm and sequencing manner , and then carrying out a new page grabbing and URL extraction [10].

At present , professional search engine network crawlers typically use the " best priority " principle to access the WEB , that is , to quickly and effectively obtain more pages related to the topic, each time the link of " most valuable " is selected for access . Because links are included in the page, and usually have higher value pages, the link also has a higher value. Unlike the general search engine, specialized search engines are limited to specific topics or specialized fields, so that the entire WEB is not required to be traversed, and only the page to be related to the subject matter needs to be selected to be accessed.

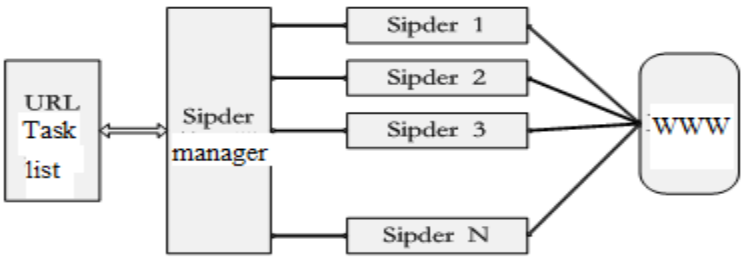

Figure1. Distributed network crawler

Cloud computing fully utilizes network and computer technology to realize the sharing and service of resources, solve complex problems such as cloud evolution, cloud control, cloud reasoning and soft computing, as is shown by figure1 and the infrastructure can be described by cloud computing architecture.

SwiftScript can be used to describe complex parallel computing based on data set type and iteration, while dynamic data set mapping can also be performed on large - scale data of different data formats. The runtime system provides an efficient workflow engine for scheduling and load balancing, which can also interact with resource management systems such as PBS and Condor to complete the task execution.

Summary

This paper presents analysis and research of distributed network crawler based on cloud computing Hadoop platform. Web crawler, flexible, in the theme of the site as a comprehensive crawling information, and can automatically construct URL recursively calls itself, multi-threaded opening fast crawling, accurate extraction of effective information stored in the database, the network delay, and be able to open the HTTPCHACHE, greatly improve the crawling speed, forge proxy the information form, let the target web site that you are safe, the integration of information.

Acknowledgements

This paper is supported by Henan key Laboratory for Big Data Processing & Analytics of Electronic Commerce, and also supported by the science and technology research major project of Henan province Education Department (17B520026).

References

[1] TianjunFu,AhmedAbbasi, HsinchunChen. A focused crawler for Dark Web forums. J. Am. Soc. Inf. Sci.,2010,6,16.
[2] Dorling. A. , SPICE: Software Process Improvement and Capability dEtermination. Software Quality Journal,2014: 209-224.
[3] Mohsen Jamali, Gholamreza Haffari, Martin Ester. Modeling the temporal dynamics of social rating networks using bidirectional effects of social relations and rating patterns, 2011: 527-536.
[4] Mandl-Striegnitz P, Lichter H. A Case Study on Project Management in Industry – Experiences and Conclusions. Proceedings of the European Software Measurement Conference (FESMA), 06. - 08. May 1998, Antwerp, Belgium, 2013: 305-313.
[5] Punam Bedi,Anjali Thukral,Hema Banati,Abhishek Behl,Varun Mendiratta. A Multi-Threaded Semantic Focused Crawler. Journal of Computer Science and Technology,2012,2,16.
[6] Ye Yunming et al. Research on distributed Web Crawler: structure, algorithm and strategy. Electronics Journal, 2003:100-108.
[7] Hongsheng Xu, Ruiling Zhang. Novel Approach of Semantic Annotation by Fuzzy Ontology based on Variable Precision Rough Set and Concept Lattice, International Journal of Hybrid Information Technology Vol.9, No.4 (2016), pp. 25-40.
[8] Gireesh Kumar P, Active Server Pages: Technology for Creating Dynamic Web Pages and Web enabled Database, Documentation Research and Training Centre, 26th-28th February, 2001:0652-2569.
[9] H.-s. XU, R.-l. ZHANG, “Semantic Annotation of Ontology by Using Rough Concept Lattice Isomorphic Model”, International Journal of Hybrid Information Technology, Vol.8, No.2, 2015, pp.93-108.
[10] Subhendu kumar pani, Deepak Mohapatra, Bikram Keshari Ratha. Integration of Web mining and web crawler: Relevance and State of Art. International Journal on Computer Science and Engineering,2010,772.
[11] Peng L, Shizhao N, Zheng W, Ziwei J, Jianwu Y, Zhongxiang Q, Wangmo P. Predicting durations of online collective actions based on Peaks’ heights [J]. Communications in Nonlinear Science and Numerical Simulation. 2018, 55: 338-354.