Research on Web Data Mining Based on Topic Crawler

Hongjian Guo

Nanjing Audit University, Jiangsu, China
E-mail: tluusdhm835155@163.com

Received 01 April 2021; Accepted 11 May 2021;
Publication 24 June 2021

Abstract

This paper analyzes the method of Web information data mining based on topic crawler. This paper puts forward the architecture of Web information search and data mining, and introduces the key technology and operation principle of the architecture. After analyzing the functions and shortcomings of ordinary crawler, this paper focuses on the working principle, implementation method and performance analysis of this crawler, as well as the functions of this crawler different from other crawlers and its application in Web information search and data mining system. The experimental results show that the crawler can get all kinds of information resources on the world wide web, which is helpful to the monitoring and management of network cultural content.

Keywords: Topic network, crawler, data mining, web information search.

1 Introduction

With the rapid development of Internet technology and the increasing popularity of the network, the network culture with digital content as the symbol and Internet as the main carrier has increasingly become one of the main

Journal of Web Engineering, Vol. 20_4, 1193–1206.
doi: 10.13052/jwe1540-9589.20411

forms of cultural communication [1, 2]. The global and open characteristics of the Internet objectively provide a hotbed for the spread of all kinds of bad network cultural content at home and abroad, and bring severe challenges to the security of China's network culture [3]. How to take high-tech means to effectively audit and monitor the content of network culture, find harmful information on the Internet in time, and prevent the erosion of bad network culture information, has become one of the urgent scientific and technological problems to be solved. Web information search and data mining is an important way to solve this problem.

2 Architecture of Web Information Search and Data Mining

The main purpose of Web information search is to find web information resources, that is, to use a technology called web crawler (or robot) to roam the World Wide Web automatically, to find new content as much as possible according to its search target, and to index and store the information searched and updated into the database [4–6].

Web mining is the application of traditional data mining ideas and methods to the web, extracting interesting, potential, useful patterns and hidden information from web resources and web activities [7]. The mined information can be used for information management, decision support and process control, and also for data maintenance.

As mentioned above, because Web information resource is a huge, widely distributed, highly heterogeneous, semi-structured and dynamic information warehouse, the traditional data mining technology can not be directly used in Web information resource mining. It needs new data model, architecture and algorithm. In this regard, we propose a web information search and data mining architecture, as shown in Figure 1.

As can be seen from Figure 1, web crawler plays an important role in the whole information search and mining system. It is the source of Internet data, which determines whether the content of the whole system is rich and whether the information can be updated in time. At present, the mainstream web information search object is still a large number of text resources with mature technology. Different from literature [3–9], we combine web content analysis with web link analysis to collect all web data including multimedia information resources such as images, sounds, pictures and even video clips, which greatly improves the quality of Web Information Mining and lays a solid foundation for the whole search and mining system.

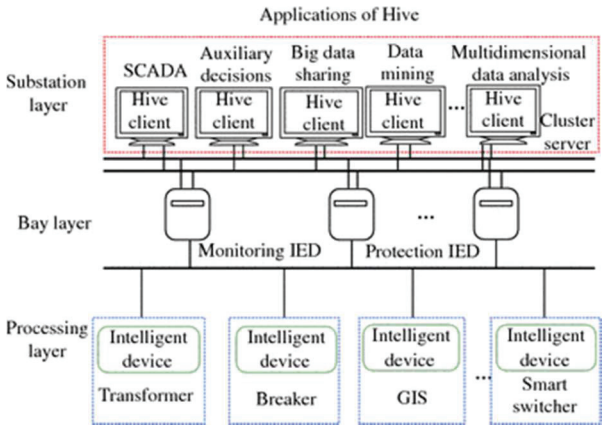


Figure 1 The search and data mining architecture of Web Information.

The search objects of Web information resources include multimedia information (including image, audio, video and other media types) and unstructured or semi-structured information (including text, HTML and other formats) [8–10]. After processing, they should be stored in the corresponding web database. Web database is composed of media database, feature database and full-text information database. Media library includes image, video, audio and so on. The feature library includes the standard description of automatic feature extraction and the feature input manually [11]. The full-text information base is used to store the source code and page URL of the obtained web page.

Because Web database can't capture and analyze business data in real time, it's not suitable for the application of DSS. Therefore, it must be exported as a data warehouse. With the help of OLAP and data mining technology, decision support system makes statistics, analysis and reasoning on the data of data warehouse to discover the potential information and rules in the data, so as to provide decision support for the management.

3 Internet Worm

3.1 The Principle of this Reptile

The world wide web is a network structure information space, which can be represented by a directed graph G = (N, E): the content of the web page is regarded as a node and uniquely marked by the URL; the links in

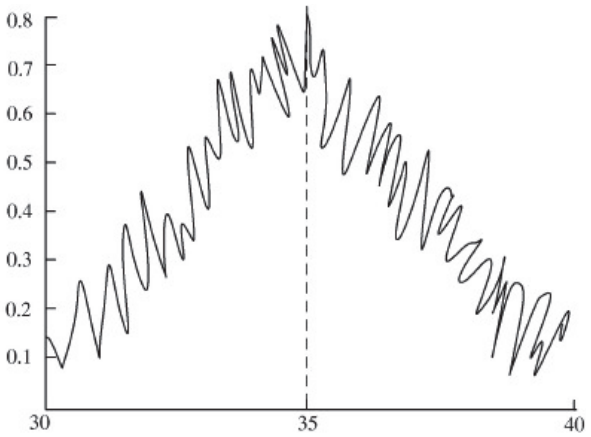


Figure 2 Representation of Web digraph.

the web page are regarded as directed edges, as shown in Figure 2. Where node set n = {N0, Nm}, and E is the hyperlink set. Leaf nodes can be web files, graphics, audio and other media files. All non leaf nodes are web files. Therefore, when crawling web pages, crawlers can use directed graph traversal algorithm (depth first algorithm and breadth first algorithm) to traverse them.

(1) Search strategy of crawler

At present, when crawling web pages, crawlers generally adopt two strategies: breadth first and depth first.

The so-called breadth first is that the crawler traverses along the width direction of the tree until it has grasped all the linked pages in the starting page, and then selects one of the linked pages to continue the process [12]. This method can make the crawler run in parallel and improve the speed of crawling.

Depth first means that the crawler traverses the unknown nodes along the depth of the tree. Because depth first is a recursive process, the crawler program will consume a lot of computer memory resources when it is executed, which will lead to the crawler into problems or even crash in many cases. In addition, recursion is not compatible with multithreading (because multithreading allows multiple tasks to run at a time, and each parallel thread has its own stack). When a method calls itself, they need to use the same stack). This crawler uses breadth first traversal strategy.

(2) Application of Multithread

In order to improve the performance of crawling, the system realizes several crawlers to access the web server in parallel, and then excavates the web page content after the connection is established successfully. Java multithreading in the specific implementation, using a thread corresponding to a URL connection. According to the actual situation, the number of concurrent threads is 13. The optimal number of threads is related to the local CPU performance and network bandwidth. The more threads on the server, the better. Therefore, we need to find a balance between the normal operation of the background database server and the rapid collection of web pages.

(3) Processing of dynamic web pages

Dynamic web pages are generated automatically from the server database by passing parameters. At present, most of the crawlers do not support the capture of dynamic web pages, especially for the web pages with "?" and "&" (and other similar symbols) in the address. We define a standard interface by analyzing the types of dynamic web pages. When crawling a web page, the crawler first determines the URL type extracted. If the dynamic web page generated by the web server meets the defined standard interface, the crawler will download the web page and establish an index database.

(4) Update of web page

In order to ensure the synchronization of database information and web content, web database needs to be updated regularly. The update of database is realized by crawler searching web space again. The specific period is how long, depending on the site update cycle. When updating the content of the website, the crawler does not have to re grab all the information content of the site, but judge whether its properties (file name and size of the document) change.

3.2 Crawler Implementation

When this crawler visits a site, it will first check whether Robots.txt (belonging to Robots protocol) exists in the root directory of the site. If it is found, the crawler will determine the access scope according to the content in the file; If the file does not exist, the crawler crawls along the link. Figure 3 describes the basic workflow of this crawler.

The crawler processing object is the URL queue to be accessed, which is initialized by the seed URL provided by users or other programs. The crawler

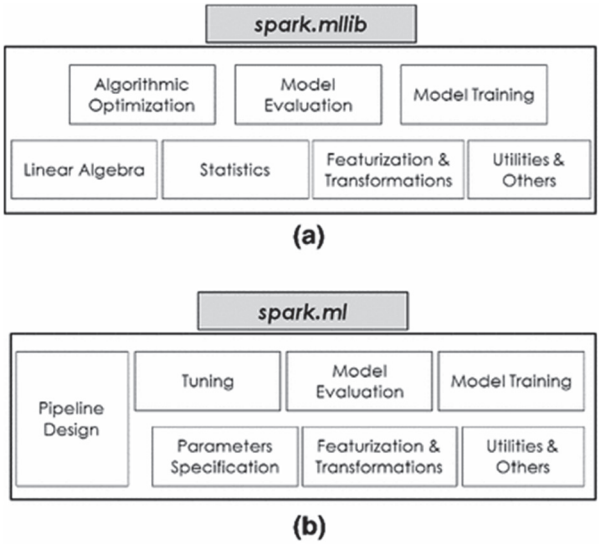


Figure 3 Basic process of web crawler.

accesses the page corresponding to the URL in the initialization URL queue through HTTP protocol, parses the web page to extract all the URLs on this page, and saves all the data on this page according to the extracted URL. Each crawling cycle selects a URL from the parsed URLs to crawl until all the URLs on the web page are crawled. Next, we will describe in detail the steps involved in the basic workflow of this crawler.

(1) Initialize URL

The pending URL queue can be constructed as a first in first out (FIFO) queue. The next page to be crawled comes from the head of the queue, and the new URL is added to the end of the queue. Each step is to select the next URL from the queue head for crawler to crawl until all the URLs in the queue are crawled.

(2) Read page

When reading a page, the crawler must first determine the type of the file. If it is multimedia data, it will be downloaded and stored directly in the web database. If it is unstructured free text or semi-structured HTML (including text, HTML and other formats), it will continue to analyze. When reading the page corresponding to the URL, if there is a timeout phenomenon, it is

considered that the page is invalid, and the corresponding URL is added to the error queue. On the contrary, if it doesn't time out, it reads the page parsing content.

(3) Analysis of web pages

After obtaining the page, we need to extract the required information by parsing the content, and guide the crawler's future crawling path. Because HTML file is composed of "text" and various "tags", the parsing process is to analyze the whole content of the source file to extract URL tag process. During document analysis, this crawler mainly extracts the following information:

(1) Document title: it is obtained by extracting the string between the identifier command <title> and </title> in the file header. Generally, the title of a web page reflects the main content and nature of the page.
(2) Link: obtained by extracting the quoted string in the identifier command < a href = >.
(3) Picture: it is obtained by extracting the quotation marks in the identifier commands < img SRC = "> and < bodybackground =" >.
(4) Video or audio: it is obtained by extracting the quoted string in the identifier command < object embedsrc = "> or < object param name =" SRC "value =" >.
(5) Multi window page: it is obtained by extracting the quoted string in the identifier command < frame or iframesrc = "> and is usually chained to the dynamic page.
(6) Base address: obtained by extracting the string of quotation marks in the identifier command.

This crawler web page analysis module will first extract hyperlinks from the above five places in HTML and do corresponding processing. For the first four tags, it will be directly taken out as the URL; for base, if the HTML defines the base tag in the header, it will take out the URL behind it as the base URL of the web page to replace the default base URL of the web page. If there is no base tag, the default web page is used as the base address.

Take the extraction of anchor tag information as an example to illustrate how the crawler extracts the URL in the document (for other images, sounds and videos, this process is basically similar). For < a href = "address" > < A/>, the "address" here is the URL address of the link target. The link target can be its own computer or the resources of any host on the Internet, such as HTML file, graphics file, text file, sound file, video file, etc. For the

extracted URL (if the URL is relative, it needs to be converted to absolute URL according to the base URL). First, you need to identify the type (because the connection on the page may point to a page, or to a picture or even an application or a compressed file. For this reason, the crawler should be able to identify the connection type to handle them accordingly. If the target URL is image, sound, video, etc., the download function will be called to store in the database directly. If the target is a page, the corresponding multithreading is enabled to read the page. If it does not time out, the page is parsed and its content is saved. If it times out, the URL is added to the wrong URL queue. Repeat the above process for the parsed URL until the target URL page does not contain other URLs.

(4) Program implementation

This crawler program uses non recursive way to realize the crawling process. When the program is implemented, four queues should be constructed: waiting queue, running queue, completion queue and error queue. Among them, the waiting queue is the collection of the initial URL of the crawler and the newly discovered URL of the crawler. The run queue is a collection of URLs that the crawler is processing. The completion queue is a collection of URLs that have been crawled. The error queue is a collection of URLs where crawlers make errors in parsing pages or when reading data. When the program is executing, a URL can only be in one queue at the same time. We call it a URL state. The program changes from one state to another according to the state diagram, as shown in Figure 4.

There are four URL state processes for a URL to be processed: first, in the waiting queue, the URL is waiting to be processed by robot, and the newly found URL is added to the queue. When robot starts to process the URL of a web page, the URL is sent to the run queue for processing. At this time, if robot makes an error when capturing a webpage, the URL of the webpage will be sent to the error queue, and the URL in the error queue cannot be moved to other queues; if robot successfully acquires a webpage, the URL of the webpage will be sent to the completion queue, and the URL in the completion queue cannot be moved to other queues. Among them, in the process of handing over the URL in the waiting queue to the running queue, we should first compare it with the URL in the completion queue to avoid repeated crawling; when a URL in the running queue is processed, the URL in the waiting queue is added to the queue according to the principle of first in first out, and the corresponding URL in the queue is deleted.

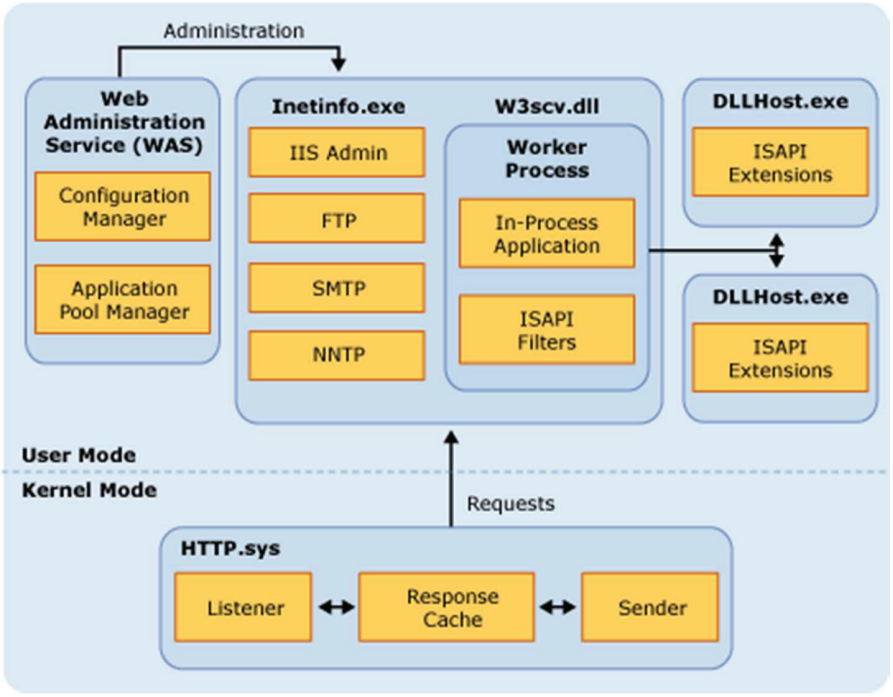


Figure 4 URL state change process.

3.3 Other Technical Problems to be Solved

The URL connection on the web page may be endless, especially when the page contains the connection of other sites:

(1) Eliminate duplication

Sometimes the pages on the site are connected to each other to form a loop. When the crawler encounters this situation, it will repeatedly execute in this loop, resulting in a dead loop. In order to avoid this situation, it is necessary to eliminate duplicate processing. Therefore, this crawler sets up two queues, one is used to store the URL that has been visited, and the other is used to store the URL queue that will be visited. When extracting a URL from a page, first of all, check whether the visited URL queue contains the URL. If it does, no processing will be performed. If it does not, it will be added to the waiting URL queue for processing.

(2) Limits

Crawlers often encounter problems such as encrypted data or web page permissions when visiting web pages. Some encrypted data crawlers can't catch them, and some web pages can only be accessed with member permissions. Crawlers can only catch these web pages after obtaining corresponding permissions.

(3) Limit depth

Some nodes on the web are very deep, so it is unrealistic to crawl such a site completely, so it is necessary to set the crawling depth for the crawler. Every time you enter the next link, it means that the depth is increased by 1. When you reach the specified threshold depth, the crawler stops searching. Generally, the depth of the site does not exceed 6. The value set in this paper is 5.

(4) Invalid link

For various reasons (reading data timeout, etc.) can not get the file or read page error, should record the invalid link URL and add to the error queue. Invalid link checking is a necessary process, which can ensure that the actual online rate of searched files or web pages is high.

4 Interface Diagram and Crawler Performance Analysis

The system is completed in the hardware and software environment of the laboratory, and the basic situation is shown in Table 1.

Regardless of the network bandwidth of the collection machine and web site and the performance and load difference of the web server, the task amount of the collection process can be roughly estimated by the number

Table 1 Software and hardware environment of crawler testing

Hardware Environment	CPU	Network DRAM	Card	Switch
	Pentium 4CPU 3.0GHZ	512M	Realtek RTL8139/810x Family Fast Ethernet NIC	10/100Fast Ethernet Switch
Software environment	Java Version	Java VM	Operating system	Development tool
	1.5	JDK	Window XP	JCreator

Table 2 Test results of reptiles

Time (Hours)	Number of Web Pages (Pages)	Average Page Download Rate (page/sec)
12	139536	323

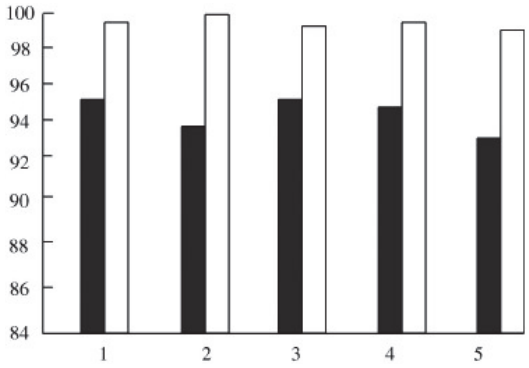


Figure 5 The test results.

of downloaded Web pages, or its performance can be reflected by testing the download rate of the crawler. For crawler, its coverage is related to the distribution of the whole web information resources, which is difficult to get in the real test. So we choose www.ccnu.edu.cn And download rate as the test object. The test was carried out in the daytime, running for one day, and the following test data were obtained, as shown in Table 2.

Table 2 shows how the entire crawler works. If we want to improve the download speed of crawler, we can increase the number of parallel crawlers.

The main reasons for not downloading all the web pages in www.ccnu.edu.cn include reading the data of web pages over time, being prohibited by Robots protocol, and being discarded because the link depth exceeds 5. As shown in Figure 5, in the process of crawling, the performance of the crawler will decrease as time goes by. The reasons include: Java language itself occupies system resources; The algorithm of crawler needs to be further optimized.

5 Conclusion

In this paper, crawler technology is used to collect and mine the resource information on the world wide web, and the whole web is integrated into a data warehouse to provide perfect web content sampling, auditing and

monitoring. In the follow-up processing, online analytical processing and data mining technology including web language analysis and logical reasoning technology can be used to analyze the data of data warehouse, and the information about web content and the access behavior and mode of group users can be obtained. Based on this, this paper evaluates the website and forecasts the police situation, so as to provide reliable and effective means for ensuring the security of network culture.

Acknowledgements

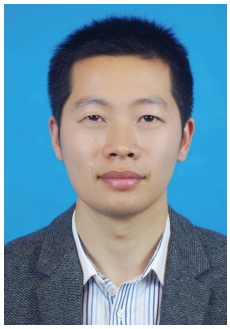
This research was supported by National Natural Science Foundation of China (Grant No. 72074117, No. 61473157) and University Natural Science Research Project of Jiangsu (Grant No. 20KJB630012).

References

[1] Yin Xiaogen, Zhang Xiaofang, Zhang Weichao. Study on 3d Reconstruction Method Based on Optical Field Digital Focusing. Photoelectron. Laser, 2015 (26): 991.
[4] Xia Zhengde, Song Na, Liu Bin. Dense Light Field Reconstruction Algorithm Based on Dictionary Learning. Acta Physica Sinica, 2020, V.69 (06): 63–71.
[5] Wang Yizhi, Zhang Xudong, Xiong Wei. Optical Field Camera Vision Measurement Error Analysis. Acta Photonica Sinica, 2017, 046(011): 113–124.
[6] Long C, Li G, Hongxing Y. 3d Dynamic Object Reconstruction Technology Based on Light Field Rendering. Journal of the University of Chinese Academy of Sciences, 2009, 26(6): 781–788.
[7] Yang Fan, Yuan Yan, Zhou Zhiliang. Study on Evaluation Method of Optical Field Camera Imaging Quality. Modern Electronic Technology, 2011, 191(1): 146–156.
[8] Liu Yanlei, Yuan Libo. Multi Directional Fourier Contour Recognition Method for Steep Edge of Objects. 2013, 7:2(2): 729–734.
[9] Tang Yi, Liu Weining, Sun Dihua. Application of Improved Time Series Model in Expressway Short-term Traffic Flow Prediction. Computer Application Research, 2015, 32(1): 146–149.
[10] Wan Ying, Han Yi, Lu Hanqing. Discussion on moving target detection algorithm. Computer Simulation, 2006, 023(010): 221–226.

[11] Liu ya, AI Haizhou, Xu Guangyou. A moving target detection and tracking algorithm based on background model. Information and Control, 2002, 12: 14–19.
[12] Pan Quan, Cheng Yongmei, Du Yajuan. Discrete moment invariant algorithm and its application in target recognition. Acta Sinica Sinica, 2001, 23(001): 30–36.

Biography



Hongjian Guo is a college teacher of Nanjing Audit University, CHINA. He attended University of NUAA where he received his B.Sc. in Computer Engineering in 2003. His work centers on Big Data and Data Mining. As a computer audit expert, he has developed many audit information systems to support data process.