Web Page Data Collection Based on Multithread

Wentao Liu
School of Mathematic and Computer Science
Wuhan Polytechnic University
Wuhan Hubei Province 430023, China
uddisoap@gmail.com

***Abstract*—The web data collection is the process of collecting the semi-structured, large-scale and redundant data which include web content, web structure and web usage in the web by the crawler and it is often used for the information extraction, information retrieval, search engine and web data mining. In this paper, the web data collection principle is introduced and some related topics are discussed such as page download, coding problem, updated strategy, static and dynamic page. The multithread technology is described and multithread mode for the web data collection is proposed. The web data collection with multithread can get better resource utilization, better average response time and better performance.**

Keywords-web page; data collection;multithread

I. INTRODUCTION

The rapid explosive growth and popularity of the web has resulted in a huge amount of information data on the internet. Many kinds of search engines are often used in order to get more useful information from the original data with the heterogeneity and the lack of structure. The first step of the process is to collect data from the complex web data which includes billions of information. The web page data collections get web information according to the web page link relationship automatically and it is often called as Robot, Spider or Crawler [1-2]. The web data includes web content, web structure and web usage. There are many problems with the web page data collections such as different protocols, different network, timeliness requirements, and diverse types of data, number of web page and web site and page quality requirements. The content data information can be divided into structured data, semi-structured data and unstructured data. The typical structured data is relational database. The unstructured data includes plain text articles, sentences, query strings, picture, sound and video. The unstructured data can't provide the useful informal directly and can't be retrieved easily. The semi-structured data includes html documents, query logs, web search results and so on. The data collection is often used for the search engine, information retrieval, data extraction and data mining. The web is noisy and redundant. A web page typically contains a mixture of many kinds of information, such as main contents, advertisements, navigation panels, copyright notices and so on. The search engine is to collect process and organize information and provide the query services. Text classification algorithm includes Naïve Bayes, SVM, KNN, Decision Tree, Neural Networks and Ensemble learning. Text clustering algorithm includes partition clustering and hierarchical clustering. The hierarchical clustering contains agglomerative clustering and divisive clustering. The information extraction is the process of automatic extraction of structured information from unstructured documents. The goal is to make information more accessible to people or more readable for machine [3-4].

The web mining discovers and extracts information from web documents and services by use of data mining techniques [5-7]. The web mining is different information retrieval or information extraction. The information retrieval is the automatic retrieval of all relevant documents and information extraction transform a collection of documents into information which is more readily digested [8-9]. The classification of web mining techniques includes web content mining, web structure mining and web usage mining.

The web structure mining generates structural summary about the web site and web page according to the hyperlink and discovers the web page structure. Web structure data describes the organization of the content. The intra page structure information contains the arrangement of html tags within a given page. The inter page structure information is hyper links connecting one page to another page. Web structure mining finds information about the web pages and retrieves information about the relevance and the quality of the web page and finds the authoritative on the topic and content. The web page includes hyperlinks which contain many annotations and it identifier endorsement of the other web page. Web structure mining discovers structure of the web and the techniques include page rank and CLEVER. The web structure mining creates a model of the web organization and may be combined with content mining to more effectively retrieve important pages. The page rank prioritizes pages returned from search according to web structure.

The web usage mining discovers user navigation patterns from the web and predicate users behavior and helps to improve huge collection of resources. The goal of web usage mining is to analyze the behavioral patterns and profiles of users interacting with a web site. The discovered patterns are usually represented as collections of pages, information, or resources that are frequently accessed by network users with same interests. The usage mining techniques includes data collection, data selection, data cleaning and data mining. The data cleaning is to remove irrelevant and erroneous references and add missing references due to caching. The applications of web usage mining include personalization

Published by Atlantis Press, Paris, France.

and it can improve structure of web site pages. It can improve caching and predication of future page references. It can improve design of individual pages and improve effectiveness of e-commerce.

The web content mining discovers using information from content of millions of sources from the web. The search engines have crawlers to search the web and gather information, indexing techniques to store the information, and query processing support to provide information to the users. The web content mining is related to the data mining and text mining. The text mining refers to data mining using text documents as data and it often uses information retrieval methods to preprocess text documents which are different from traditional data preprocessing methods used for relational tables. There are many data mining techniques can be used in web content mining and many web contents are texts. In traditional data mining the data is structured and has relational well-defined tables, columns, rows, keys, and constraints. In web mining the web data is semi-structured and unstructured and readily available data rich in features and patterns. The web content mining techniques include classification, clustering and association. The documents classification is supervised learning which is a machine learning technique and make a function from training data set. The documents are categorized and the output many predict a class label of the input object. The classification techniques include nearest neighbor classification, feature selection and decision tree. The feature selection removes terms in the training documents which are statistically uncorrelated with the class labels. The document clustering evolves measures of similarity to cluster a collection of documents into groups.

The web has become an important way for people to obtain information. But at the same time the internet is an open dynamic and globally distributed heterogeneous network. The network resource distribution is very fragmented and there is no unified management structure which leads to difficulties of access to information. How to quickly and accurately find the required information from the vast information resources has become a major problem for the network users. Since this information exists in the form of semi-structured or free text disperse a large number of web pages and it is difficult to directly access and use. Information collection for the site will be able to play a significant role. It not only can be positioned directly to the information required by the user, and uses a certain way to increase the semantics of these information and mode information. The network data collection means by automatically obtain the link relationship between the web page from the web page information, and with the link to keep the desired web page expansion process. In this process, the web crawler find page from the web page and download. The entire web network can be seen as a directed graph. The crawler starts from seeds URL with the breadth first style and save it to the local file. The crawler parses out the content of the pages that contains the URL link and adds the URL to the URL collections. This process is repeated until the URL collections of all links have been collected over. Acquisition time has already reached the required or all of the connection does not exceed a certain depth has been collected over. The downloaded pages standard html text, as well as the acquisition URL, acquisition time elements must be recorded. Web information extraction is to extract information of interest to the user, no structure or semi-structured HTML pages and converting it to structured data stored. Web mining is extracted from web documents, web activities of interest, potentially useful patterns and hidden information. Web mining can play a role in many aspects, such as the structure of the search engine excavation to determine the authoritative pages, web document classification, web log mining, intelligent query create web metadata warehouse and so on. Web content mining refers to summarize large document collections on the web, classification, clustering, association analysis, and web document trends. Web content mining makes the data mining technology in the network information processing. Web content mining is different from the traditional data mining techniques. Web mining is mainly for a variety of unstructured and semi-structured data, such as text data, audio data, video data, graphics, image data of a variety of data such as the integration of multimedia data mining. Web mining can be divided into two kinds of mining based on text mining based multimedia. Information and knowledge that can be extracted from the link between the ultra documents. He can find the web pages of the influential and authoritative and found that the structure of the web documents. He can find hierarchical network structure found in the special field of website. With the rapid development of computer and internet technology a sharp increase in the amount of information on the network. It has become the human history, the most number of resources, resource types and most complete, a comprehensive information resource largest library. Rich source of information, widely distributed, and the types of information resources in heterogeneous distributed in cyberspace. How accurate and effective access to information on the internet has become a daunting task, solve this problem the best solution is to use the search engine. Web document hyperlink can be recursive access to the new page. Its main function is to automatically crawl the web documents from web sites on the internet and extract some information from web documents to describe the web documents. It adds and updates data for search engine site database server. These data include the various links in the HTML, the title, the length, the file creation time, HTML files, etc. In this era of rapid development of information technology, the information acquisition, processing and applications have become the key to the development of economic, scientific, military, cultural and other fields activities. The access to information is the beginning of the three-step, has an especially important role in the field of information technology. Internet is an incredible speed in rapid development, it accommodates vast amounts of various types of raw information, including text messages, voice information, image information, and so on. How to master the most effective heterogeneous, unstructured text of web information is a major goal of information processing. An emerging technology in the field of natural language is the information extraction technology for unstructured data

information mining. The technology extract, filtering irrelevant information, the text information to the user in the form of concern to the re-organization of efficient reorganization. Information extraction is loosely structured natural language information extraction into structured semantics explicit form of using the computer for the efficient storage and use. Web text mining is to find effective, innovative and potentially useful and ultimately understandable knowledge from large amount of unstructured, heterogeneous collection of web documents.

II. WEB PAGE DATA COLLECTION

The rapid growth of the World Wide Web exceeded all expectations in recent years. There are several billions of HTML documents, free txt content, unstructured information, videos, pictures, sounds and other multimedia files available via internet and the number is still rising. How to retrieve interesting content has become a very difficult task because of the immense web content due to heterogeneity and lack of structure of web data. The crawlers collect web page information for the web mining or search engine. The crawler traverses the hypertext structure in the web and collects information from visited pages. The traditional crawler visits entire web and get information from the web content. The periodic crawler visits portions of the web and updates subset of index. The incremental crawler selectively searches the web and incrementally modifies index. The focused crawler visits pages related to a particular subject. The focused crawler only visit links from a page if that page is determined to be relevant. The classifier assigns relevance score to each page based on crawl topic. The web content mining is the process of extracting knowledge from web contents. Web crawler from one or more of the initial page URL, URL list on the initial page; crawl pages, and continuously extracted from the current page, the new URL into the pending crawling queue until stop condition.

The depth-first algorithm from the start page, along on a link has been search up to a file that does not contain any links to form a complete chain and then return to continue to choose other links to similar visit. The sign of the end is no other hyperlinks can search. The advantage of this algorithm is theoretically capable of traversing deeply nested pages in a web site until it encounters the depth of the search tree. Breadth-first algorithm is to finish searching all the links in a Web page and then continue to the next level search, until the bottom. It overcomes the completeness and optimality shortcomings do not have the depth-first algorithm. It can reduce to the same server has been accessed frequency, but have a larger time complexity and space complexity. Some studies have shown the importance of a better breadth-first approach is the collection of pages. Collected regularly refers to the re-acquisition of all pages after a period of time to replace the original pages, all mined out. Incremental acquisition is only collected according to some strategies that may add, change pages, and delete pages that are no longer exists. The network information collection is a data extraction from a process of landing pages removal of some of the data to form a unified local database. These data have been only visible pages in text form exists only for people to read and cannot be processed. A complex data extraction process need to deal with obstacles, such as the session identifier, HTML forms, the client JavaScript, and data integration issues such as missing data and repeat. Network information gathering is unstructured information extracted from a large number of web pages saved to the process in a structured database. How effective mining the network information successfully, how to gather information outside the enterprise, is essential for the company's business. System to collect useful data for a specific site and the data is stored as a text file or data format and can analyze, and process and filter data to obtain valuable data.

The process of getting html content from one URL is shown in Figure 1.

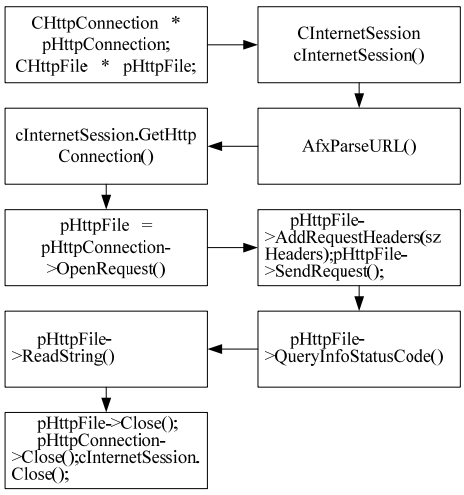

Figure 1. The process of getting html content

Web collection system must comply with the limit posted on the website of Robot.txt acquisition agreements, the acquisition time try not to be too dense collection of a website, such intensive visit is similar to DoS attacks, resulting in normal ordinary users browse the site generates difficulties. Some sites will be tightly controlled intensive access behavior. The pretreatment of the raw data format analysis, conversion and language recognition, coding recognition and conversion, for example, GB, BIG5, Unicode, UTF-8 between the conversion. The classification clustering search engine will retrieve prior classification, can achieve more accurate search, reduce the consumption of retrieval, but also easy to search results organization and display. This process is starting from the initial URL set, all of these URL into an orderly queue to be collected. Collector from the queue in order remove URL, through an agreement on the Web, get the URL points to the page, and then extracted from these pages get a new URL, and they continue to put into the queue to be collected , and then repeat the process until the collector according to their own strategy to stop the acquisition. Web seed queue storage interested website URL, it can be manually added by web logger found in the collection process. Collection list keep the visited URL link. The dynamic html content can be got by POST method. The collected rules are xml format which

include main URL and boundary word for title, content, date and so on. The XML format is shown below.

```
<Rule>
<main_url>main url </main_url>
<url_start></url_start>
<url_end></url_end>
<title_start></title_start>
<title_end></title_end>
<date_start></date_start>
<date_end></date_end>
<content_start></content_start>
<content_end></content_end>
<paper_directory></paper_directory>
<start_page_number></start_page_number>
<end_page_number></end_page_number>
<utf_code></utf_code>
</Rule>
```

The main_url is the first URL for crawling and it contains more sub URL address. The url_start and url_end explain the sub URL information. The paper_directory is used for saving the html file. The start_page_number and end_page_number are used for getting URL in bulk because one topic html may be has more sub pages. The utf_code describe the code of HTML whether it is UTF.

III. MULTITHREAD

The web data collection often uses the multithread to get parallelism and improve efficiency on gathering the web information. Multiple threads can be created within a same process and share code, address space, and operating resources for the process. The thread programming models include the boss worker model, the peer model and the pipeline model [10]. There are three popular thread libraries which includes pthreads, Java threads and Win threads. The thread synchronization is the key problem of using threads. Sections of code that access shared data are referred to as critical sections. Mutual exclusion or thread synchronization makes sure that a thread accessing the shared data excludes all other threads from doing at the same time. Several thread synchronization mechanisms can be applied in the multi thread programming. There are many methods to make synchronization in win multithread programming and it includes interlocked functions, critical sections, wait functions, mutexes, semaphores and events.

The multithread mode of getting URL is shown in Figure 2. The monitor thread is constantly listening the unvisited URL queue which stores the seed URL and unvisited URL from the parsing html thread. The monitor thread removes a URL from the unvisited URL queue and makes a parsing html thread which can analyze the whole html page and get sub URL or other useful information and add the unvisited URL to the unvisited queue and add the visited URL to the visited URL list. If the URL is exists in the visited URL list, the URL is discarded. Every URL has one data structure which contains the depth of the URL node in the whole URL tree. The depth of the seed URL node is zero. When the maximum depth of one URL is reached, the URL is abandoned. When the unvisited URL queue is empty, the monitor thread can be exited.

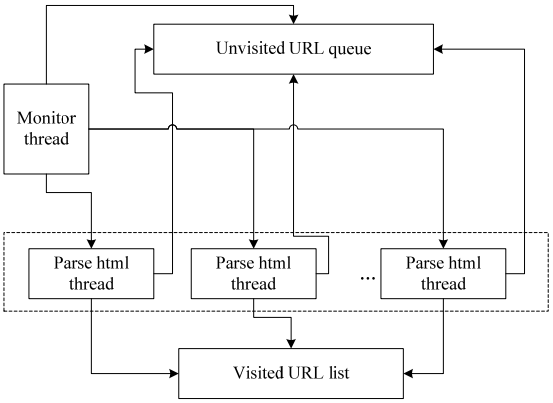

Figure 2. Multithread model

IV. CONCLUSION

The web data collection is often used in the information extraction, information retrieval, web data mining and search engine. The web data collection with multithread can get better resource utilization, better average response time and better performance. Multithread enables to write efficient programs that make the maximum use of the CPU, keeping the idle time to a minimum because the transmission rate of data over a network is much slower than the rate at which the computer can process it. The research trends include high-speed and high-quality information collection, personalized information collection and topic-based information collection.

REFERENCES

[1] Christopher Olston and Marc Najork,Web Crawling,Foundations and Trends in Information Retrieval,2010

[2] Carlos Castillo and Ricardo Baeza-Yates. A new crawling model. In Poster proceedings ofthe eleventh conference on World Wide Web, Honolulu, Hawaii, USA, May 2002.

[3] Soderland, S.,Learning information extraction rules for semi-structured and free text. Journal of Machine Learning, 34(1-3): 233-272, 1999

[4] Freitag, D., Information extraction from HTML: Application of a general learning approach. Proceedings of the Fifteenth Conference on Artificial Intelligence,1998.

[5] Bing Liu., Web Data Mining: Exploring Hyperlinks, Contents and Usage Data, Springer, 2009.

[6] Baldi, P., Frasconi, P., & Smyth, P. Modeling the Internet and the Web. Probabilistic Methods and Algorithms. Chichester, UK: John Wiley & Sons. 2003.

[7] Jiawei Han and Micheline Kamber. Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers, 2006.

[8] Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, Introduction to Information Retrieval, Cambridge University Press, 2008.

[9] R. Baeza-Yates and B. Ribeiro-Neto. Modern Information Retrieval. ACM Press, New York,1999

[10] Programming with POSIX threads, by D. Butenhof, Addison Wesley 1997.