# Analysis and Design of Improved Intelligent Search Strategy for Web Crawler

Hongsheng Xu[1,2 a *], Bin Zhao[1,2] and Ganglong Fan[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:** Intelligent search; Network crawler; Spider; URL; Link filtering

**Abstract.** This paper mainly studies the design and implementation of the search engine's searcher Spider program, and introduces the concept and technical essentials of Spider program in detail. Network crawler is a web crawler program which can run in the background with configuration file as the initial URL crawling down with the width first algorithm and saving the target URL. The paper presents analysis and design of improved intelligent search strategy for Web crawler. Based on multi-thread web crawler, the client can access the server through socket, and the client sends its own set request to the server.

## Introduction

The representative of the search engine based on the popularity of access is direct hit; its basic idea is that the website that most people choose to visit is the most important one. Ask the websites and the time they spend on them to determine the importance of the sites, therefore, this evaluation system has the same disadvantages as the search engine based on link evaluation.

With the rapid development of the Internet, the World Wide Web has become the carrier of a lot of information. How to extract and utilize this information effectively becomes a huge challenge. And Google, as a tool to assist people to retrieve information, become the portal and guide for users to access the World Wide Web. However, these universal search engines also have some limitations, such as: different fields, Users from different backgrounds often have different retrieval purposes and requirements. The results returned by a generic search engine contain a large number of pages that users do not care about [1]. The goal of a generic search engine is to achieve as much network coverage as possible. The contradiction between limited search engine server resources and infinite network data resources will be further deepened.

Focused web crawlers selectively visit Web pages and related links in accordance with established grab targets; grab the required information according to the specified rules, and its general structure. Unlike common web crawlers, The user needs to provide a topic description to specify the grab target, and in order to ensure that the acquired page is relevant to the theme, on the one hand, the page needs to be evaluated according to the relevance of the topic description and filtered out of the irrelevant page; on the other hand, Only those that are evaluated as useful will be added to the pending URL queue. Topic description, page filtering and link filtering are the focus of research on web crawlers.

Is a web crawler, according to certain rules, information capture web automatic program or script. In addition, some do not use names and ants, automatic indexing, simulation program or a worm. The network search function in Internet content brought the explosive development of content retrieval needs. Search engine development continuously [2]. People's needs are constantly improving; the network information search has become the people every day to carry out the content. How to make the search engine can always meet the needs of the people. To achieve the initial retrieval function by indexing the station, and the network robot, namely the web crawler technology, search engine era began to hair get out of hand.

The traditional crawler starts with the URL of one or more initial web pages and obtains URLs on the initial web pages. During the process of crawling the web pages, new URL are constantly extracted from the current page and put into the queue. Compared with the traditional crawler, the workflow of focused crawler is more complicated, so it is necessary to filter the topic-independent links according to a certain analysis algorithm of web pages. Keep the useful link and put it into the URL queue waiting to be fetched [3]. Then, it will select the next page URL from the queue based on a search strategy and repeat the above process until a condition of the system is reached. All crawler crawled web pages will be stored, analyzed, filtered, and indexed to facilitate subsequent query and retrieval.

The key technologies of network crawler in this paper are as follows: socket technology: http protocol technology, regular expression technology. Based on the multithread crawler model, the SOCKET correlation function is used to send requests to the source URL according to the HTTP protocol. Then the target URL is used as the source URL for a new round of downward crawling search, and then the whole Network Spider crawls down according to the width first algorithm.

## Analysis of Spider Working Principle of Network Crawler

It is almost impossible for a search engine to grab all the web pages on the Internet. The largest search engine has just grabbed about 40% of the total number of web pages. The reason for this is, on the one hand, the bottleneck of crawling technology, which can't traverse all the pages. Many pages cannot be found in links to other pages; another reason is the problem of storage and processing techniques, if calculated according to the average size of 20 K per page (including images, the capacity of 10 billion pages is 100 ×2000G bytes), Even if it can be stored, there is a problem with downloading it (at 20K per second on a single machine, 340 machines need a year of incessant download time to download all the web pages.) and because of the amount of data, Therefore, many search engine web crawlers only grab important web pages, and when crawling, the evaluation of importance is mainly based on the link depth of a web page.

With the enrichment of the data form of the World Wide Web and the continuous development of the network technology, pictures, databases, audio and video multimedia and other different data appear in large numbers [4]. General search engines are often unable to do anything about these data, which are dense in information and have a certain structure, and can not be found and obtained well. Most of the general search engines provide keyword-based retrieval. It is difficult to support the query based on semantic information, as is shown by equation (1) [5].

$$\hat{x}^{(1)}(k+1)=\left[x^{(1)}(1)-\frac{\hat{u}}{\hat{a}}\right]e^{-\hat{a}k}+\frac{\hat{u}}{\hat{a}} \tag{1}$$

In order to solve the above problem, focused crawler has emerged as the times require. Focused crawler is a program that automatically downloads web pages. Selective access to web pages on the World Wide Web and related links to get the information you need. Unlike generic crawlers, focused crawlers do not seek large coverage, but rather target pages that are relevant to a particular topic. Prepare data resources for topic-oriented user queries.

This is a special storage data structure. WebDB is used to store all site structure data and attributes from scratch (including refetching) to a collection of structures and attributes that are captured by the site data. WebDB is only used by the crawler. The search program does not use it to store two types of entities: pages and links. Pages represent a web page on the network whose Ural is indexed as a token. At the same time, MD5 hash signature is established for the content of the web page. Other content related to the page is also stored. Include: the number of links in the page (external links, page grab information (in case the page is repeatedly fetched, and a score representing page level score. Links represent links from one page to other pages). So WebDB is a network diagram, the node is the page and the link is the edge.

Using the Spider class and the "ISpiderReportable" interface, you can easily add "spider" functionality to a program. Here's how the .Spider class of the Spider class works must keep track of the URL it has accessed. The purpose of this is to ensure that the spider does not access the same URL more than once; further, the spider must divide the URL into three groups, the first of which is stored in the "workloadWaiting" attribute. Contains an unprocessed list of URL, in which the first URL to be accessed by Spider also exists; the second group is stored in "workloadProcessed", which has been processed by Spider and does not need to be accessed again; and the third group exists in "workloadError", which contains the URL where an error occurred [6].

Vertical search: also known as professional search, high-speed, massive and accurate crawler DataScraper is the strength of the problem, 24 hours a day, 7 days a week, 7 days a week, the periodic batch of unattended scheduling, Add breakpoints and software watchdog watch Dog. Mobile Internet: mobile search, mobile phone mashup.com, mobile social networking, mobile e-commerce can't do without structured data content data carper real-time and efficient content collection, the mobile Internet: mobile search, mobile phone mash pan, mobile social networks, mobile e-commerce can not do without structured data content, data carper real-time and efficient collection of content. Output grab result files in XML format rich in semantic metadata, ensure automated data integration and processing, overcome barriers to small screen display and high precision information retrieval. Mobile Internet is not a subset of Web, but all. The bridge is set up by MetaSeeker.

The crawler starts from one or several "initial URL, the initial URL on the page, in the process of crawling, continuously from the current page from the new URL queue, until the system must stop conditions. Analysis of dynamic web crawling content: analysis of dynamic web page parameters, according to certain rules, fight" all to be content to crawl URL, just grab the specific range of dynamic pages. Analysis of the special content capture content: such as RSS, XML data, special circumstances require special treatment. Such as the rolling news page news, need to constantly monitor spider scanning, discover new content immediately crawl.

## Analysis and Design of Improved Intelligent Search Strategy for Web Crawler

Spider is a crawler crawling around the web. A web crawler searches for a page by using a web page's link address, starting with a page (usually the home page), reading the contents of a page, and finding other link addresses in a web page [7]. And then look for the next page through these link addresses, and this goes on until all the pages of the site are captured. If you think of the entire Internet as a website, then the web crawler can use this principle to capture all the web pages on the Internet.

In a crawler system, an important part of the URL queue is to be fetched. The order in which the URL in the URL queue is to be fetched is also an important question, as is shown by equation (2), because it involves fetching that page first. The method that determines the order of these URL is called the grab strategy.

$$x^{(1)}(t) = \left( x_1^{(0)} - \frac{u}{a} \right) e^{-a(t-1)} + \frac{u}{a} \tag{2}$$

Search strategies based on content evaluation are based on topics such as keywords, Topic related documents) to evaluate the value of the link and determine its search strategy: link text refers to the descriptive text around the link and the text information on the link URL. Search strategy based on link structure evaluation: a method for determining the importance of links by analyzing the relationships between Web pages to determine the link access order. Pages with more links in or out of the chain are generally considered to be in or out of the chain. PageRank and Hits are representative algorithms [8].

The Fetchlist of Segment is the list of url used by the crawler. It is the output data of .Fetcher generated from WebDB. The output data of .Fetcher is retrieved from fetchlist. The output data of .Fetcher is first indexed in reverse [9]. The indexed results are then stored in the segment. Segment life cycle is limited; it won't work when the next grab starts. The default refetch interval is 30 days. So it's OK to delete segment beyond this timeframe. And you can save a lot of disk space. Segment is named date plus time. So it's very intuitive to see their life cycle.

First, you don't have to do your own analysis of Dom structures, you have ready-made libraries, and programming isn't complicated; second, you can implement very complex but flexible positioning rules, and regular expressions are hard to write. If positioning is to consider HTML file structures, it is not easy to parse HTML files with regular expressions, and it is much easier to parse Javascript files if you hand over the task to a ready-made library. 4th, assuming that you have to parse the contents of Javascript, Regular expressions are powerless, and DOM itself is powerless, but it's possible to take advantage of the power of a particular platform to extract the content of a AJAX site.

Fetching is a circular process: the grab tool generates a fetchlist collection from the WebDB, the extraction tool downloads the web page content from the network according to the fetchlist; The utility updates WebDBs according to the new links found by the extraction tool; then generates a new fetchlist. This grab loop is often referred to as the generate/fetch/update loop in nutch.

## Experiments and Analysis

Traditional web crawler technology is mainly used to capture static Web pages. With the popularity of AJAX/Web2.0, how to grab dynamic pages such as AJAX becomes an urgent problem for search engines, because AJAX subverts the traditional pure HTTP request / response protocol mechanism. If search engines still use the "crawling" mechanism, it is impossible to capture valid data from AJAX pages. AJAX uses an asynchronous request / response mechanism driven by JavaScript, and previous crawlers lacked the semantic understanding of JavaScript. Basically, it is impossible to simulate the asynchronous call that triggers the JavaScript and parse the returned asynchronous callback logic and content.

The URL task list, the initial URL seed, stores the initial seed in the critical region, and then, using multithreading, each thread sends a request, and multiple crawler threads crawl over the Internet simultaneously. Grab the web page according to a certain search algorithm and extract the URL to return to the session area, The obtained URL is sent back to the URL list as a new seed to start a new round of multithreaded search and grab web pages and extract URLs so that the whole crawler program can run in a circular way [10].

The search strategy analysis gives the crawler a starting IP address, and then searches the document in each WWW address after the IP address segment according to the IP address increments. It does not take into account the hyperlink addresses of each document to other Web sites. The advantage is that the search is comprehensive and the source of information for new documents that are not referenced by other documents can be found; the disadvantage is that it is not suitable for large-scale searches.

Based on the analysis of the use case diagram, the page collector and the page indexer are the core modules of the search engine and the core of this project. Therefore, the design of this project is based on the background running project, is a very powerful program, it will be based on the pre-set URL to query the corresponding web pages, According to the links in the web page, the web crawler's access to the page is the process of traversing the information on the web to meet the needs of the customers.

According to the HTTP protocol, the network crawler constructs the sending request data of the Sendler function. A char type array sendBuf is defined to store the request information. The concrete implementation is as follows: char sendBuf; then the request information is constructed according to the HTTP protocol. The GET request method is used to send the request. In order to respond to the negative web crawler search crawling, the general website will produce a map of the site named sitemap.jsp.

In a width first search, search for all the hyperlinks in a Web page, then continue to search the next layer until the bottom layer. For example, there are three hyperlinks in a HTML file, and select one of them to process the corresponding HTML file. Instead of selecting any hyperchains in the second HTML file, return and select the second hyperchain, process the corresponding HTML file, and return, Select the third hyperchain and process the corresponding HTML file. Once all the hyperchains on a layer have been selected, you can start searching for the remaining hyperchains in the HIML files you have just processed. This ensures the first processing of the shallow layers.

In addition, in the application of AJAX, there are a lot of changes to the DOM structure. Even all the content of the page is read directly from the server side through JavaScript and drawn dynamically. This is incomprehensible to the static page which is used to the relative invariance of the DOM structure. For technologies like AJAX, the required crawler engine must be event-driven.

## Summary

The crawler program in this paper can read out the URL task list from the configuration file, that is, the initial URL seed, and save the initial seed in the critical region. Then, using multi-thread technology, each thread sends a request. Multiple crawler threads crawl over the Internet at the same time, search and crawl web pages according to the breadth search operation and extract URL to return to the session area. The URL obtained by regular expression filtering is stored in the session area and then sent back to the URL list as a new seed. The recursive algorithm is applied to start a new round of multithreaded search and grab web pages and extract URLs, so that the whole crawler program can run circularly.

## 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] Nie J Y. A general logical Approach to inferential information retrieval. Encyclopedia of Computer Science and Technology, 2011:203-226.

[2] Saggion H, Lapalme G.Concept Identification and Presentation in the Context of Technical Text Summarization.In: Proc of the Workshop on Automatic Summarization, New Brunswick, New Jersy: Association for Compution Linguistics, 2010: 1-10.

[3] 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.

[4] Claudio Carpineto, Renato De Mori, Giovanni Romano etal. An information- theoretic approach to Automatic query expansion. ACM Transactions on Information System, 2011, 19(1): 1-27.

[5] Xu Lili. A comparative study of professional search engines, Modern Intelligence, No. 1, 2005, No. 188-190.

[6] Morris A H, Kaspcr G M, Adams D, The Effects and Limitations of Automated Text Condensing on Reading Comprehension Performance. Information Systems Research,2012, 3(1).

[7] Wu Hongqin. Study on the Classification system of Chinese search engine, Library Journal, No. 3, 2015, No. 69-70.

[8] 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.

[9] Bao Dongmei, Zhou Yueqing. Evaluation of search performance of famous Chinese and English search engines, Modern Library and Information Technology issue 1: 36-40, 2014.

[10] Turney P.Learning algorithms for key phrase extraction, Information Retrieval, 2010; 2(4): 303-336.