International Refereed Journal of Engineering and Science (IRJES)
ISSN (Online) 2319-183X, (Print) 2319-1821
Volume 6, Issue 4 (April 2017), PP.16-19

The Research on Related Technologies of Web Crawler

Wang Bingwei[1], Yu Su[2]

(College Of Mechanical Engineering, Shanghai University of Engineering Science China)

ABSTRACT: Web crawler is a computer program which can automatically download page or automation scripts, and it is an important part of the search engine. With the rapid growth of Internet, more and more network resources, search engines have been unable to meet people's need for useful information. As an important part of the search engine, web crawler is becoming more and more important role. This article mainly discusses about the working principle, classification of web crawler, etc were related in this paper. And then discusses the research and the subject of the search engine important topic web crawler.
Keywords: Web crawler, Search engine, Topic web crawler.

I. INSTRUCTION

In recent years, with the development of the Internet in the huge network information, all users are available through certain means to get their want of knowledge, including life, science and technology, and the content of the military, etc. Normally, the search engine can search for numerous of web on the net, and at the same time, it can be able to complete the index action of input keywords. In the computer technology, efficient and accurate access to the target content needs to be solved so as to create conditions for search engines. Web crawler in the search engine is the core part of it also due to the presence of the engine and begins to be attention.

Web crawler's principle of work is usually running from which called a seed set of URL, it will first of all put these URL in an orderly crawler queue. According to certain order take out URL and download from the analysis of the content of the page, extract the new URL and deposit to crawl URL in the queue. It repeats the above process until the URL queue become empty or satisfied a crawling termination conditions.

Web crawler is directly faced to Internet as one of the basic components of search engine. It is a data source of search engine determines whether the content of the whole system and the information of the whole system can be updated in a timely manner. Its performance directly affects the effect of search engine.

II. THE CLASSIFICATION OF WEB CRAWLER

Web crawler can be divide into the following several types in accordance with the system structure and implementation technologies General Purpose Web Crawler. Focused Web Crawler, Incremental Web Crawler and Deep Web Crawler. The actual web crawler system is usually accomplished by combining several crawler technologies.

2.1 General Purpose Web Crawler

General Web crawler: URL expands to the entire Web crawling object from some seeds, mainly for the portal site search engine and large Web service providers to collect data. Due to business reasons, some technical details released very little. This kind of web crawler to crawl range and a huge number, higher requirements for crawling speed and storage space, to crawl the page order requirements are relatively low, because at the same time to refresh the page is too much, usually adopt parallel works, but takes a long time t o refresh the page at a time. Although there are some defects, general web crawler is suitable for the for the search engines for a wide range of topics, has the strong application value. General web crawler structure can be roughly divided into pages crawled module, a page analysis module, link filtering module, database, URL queue, initial URL set several parts. In order to improve work efficiency, general web crawler will crawl to take certain strategy.

2.2 Focused Crawler

Focused web crawler is selectively crawling with predefined topic page web crawler. Compared with the general web crawler, focused crawler need only crawl pages on the subject, and greatly save the hardware and network resources, save the page also due to fewer updates fast, can also meet some certain people demand for information in specific areas. Focused web crawler compared with the general web crawler, increased the link evaluation module and evaluation module, focused crawler to crawl strategy implementation is the key to

evaluate the importance of the page content and link, the importance of the different methods to calculate the different, the resulting links have different access sequence.

2.3 Incremental web crawler

Incremental crawler is refers to the incremental updates on some downloaded web pages. And at the same times only crawl new or have changed the web crawler. To some extent, it can ensure the crawl the page is as far as possible the new page [16]. And Compared with the periodic crawler, incremental crawler can only crawl in need of new or updated pages, no change is not to download page, which can effectively reduce data downloads, update has crawled web pages, reduce the cost of time and space, but increased the crawling algorithm complexity and difficulty.

2.4 Deep Web Crawler

Surface Web page refers to the traditional search engine to index page, is given priority to with hyperlinks to static Web page of a Web page. Deep Web is the most content obtained by static link, not to hide behind the search form, only the user submits some key words to get a Web page. Such as the content is visible after the user registration page is Deep Web.2000 Bright Planet points: Deep Web accessible information capacity is in the Surface Web hundreds, is the Internet's largest and fastest growing new information resources.

III. CRAWLING TOPIC CRAWLER

In the process of web crawling the crawler for new web links and judge the link whether is relative to the topic. After the meet certain requirements, it puts the link into the URL of the queue which wait for download or give up the link. So it can focus on grasping the theme related pages and increase the rate of recall of the crawler, spiders crawl range and be narrowed.

The structure of the topic web crawler

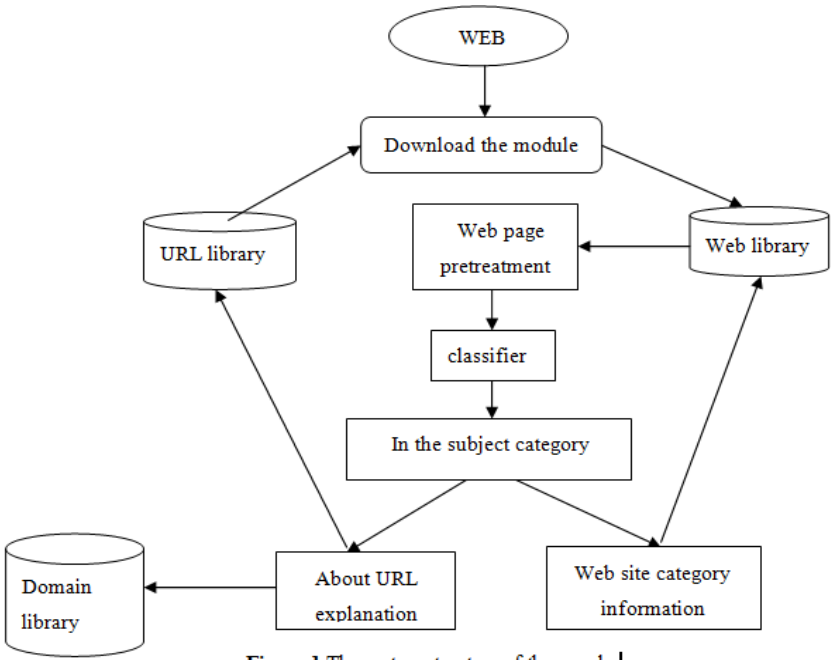

Figure.1 The system structure of the crawler

From the architecture diagram we can know that topic web crawler is divided into five parts which includes the data storage, download module, pretreatment, web page classification and link analysis.Search strategy is the core of the crawler. It largely determines the collecting efficiency of the crawler. Links and the extraction of score is the key to search strategy. Links the extraction of the need to follow certain agreement some links based on the ROBOTS protocol to cleaning. After extraction of good link score is decided by the order of the crawl which is a hotspot of current web crawler. In order to obtain higher coverage general search engine WEB crawler is often times book method is used to iterate through the WEB. It is different from the general search engines professional search engine service to specific people. Its search restricted to a particular topic or the content of the specific field. The total figure in the search process does not need to traverse and

simply select page associated with the theme of the page for a visit.

Existing network spider search strategies can be divided into two categories, they are immediately return value evaluation based search strategy and the search strategy based on future return value evaluation. Immediately return value evaluation based search strategy includes based on content evaluation and link structure evaluation based search strategy. Based on the content of evaluation of network spiders mainly evaluation of the theme and anchor text links according to the similarity. Anchor text refers to the text around the links and text similarity calculation usually use vector included Angle cosine. Such as:

$$\text{SIM}(q, p) = \frac{\sum_{k \in q \in p} w_{kq} w_{kq}}{\sqrt{\sum_{k \in p} w_{kw} p^{w_{kw}}}} \qquad (1)$$

Among them, the p collection subject keywords, q represents the page anchor text, on behalf of the k in the main collection q in words Perception of the importance of problem q, usually use the formula to calculate. Search strategy based on link structure evaluation is based on the analysis of the importance of the relationship between the page reference each other links and then decided to link access sequence. It is generally thought more page pointing or point to more web pages with high value of the page. Search strategy based on future return value evaluation, the strategy that web spiders can more accurately predict the value of a web page. Main method of cement machine learning and the search strategy based on the context diagram

3.1 The definition of topic web crawler

The crawling topic crawler is according to predefined topic, a certain analysis algorithm to crawl web topic analysis filtering and topic not related web pages. In the process of fetching the relevant web pages with the links are relevant in the crawl to the queue until it reaches a certain condition. It doesn't have to collect all of the web page. The web crawler only crawlers the page on the subject and focus on the topic of web links. And then extracted from the download page URL and predict whether the URL with the given topic. According to the priority order to access the URL will be unrelated to give up as much as possible in the process of crawling find pertinent to the topic and download page.

General web crawler page to gather as much as possible, in the process does not consider processing sequence of the page and page can be obtained is relevant to the topic. Topic WEB crawler based on traditional crawler, joined the WEB data mining and other related technology enables the crawler in the process of crawling along the path of the target page can be found for as much as possible. So as to improve the accuracy of the existing search engines and the update cycles.

3.2 The basic idea of topic web crawler

The topic web crawler is based on the basic idea that prepared. And it analyzed the theme of the already downloaded content of the page and the links in the page. Calculate the current page level on the subject and predicted that the next link which need to deal with. Ensure to get as much as possible in the process of crawling is closely relationship with theme page. In this process topic web crawler need to use a certain algorithm to filter out links that are not too big relations in the page and keeps those may be related to the theme of the close relationship between link and put it in the relevant filed. The study of the topic web crawler includes the following aspects. The description of the theme: how to describe the object of crawl. The different between topic crawler and the general crawler is the topic crawler only crawl web pages on the subject. Accurate definition and description of the subject can help topic crawler found the net on the subject more effectively.

Topic crawling strategy is lined up access sequence. In the process of scraping of the page the theme crawler will decrease correlation is passed to the child links according to certain rules. And insert the link on the subject. When crawl again the topic crawler not simply carried out in accordance with the breadth or depth of preference but according to the link to the topic of relevance to sort. URL to crawl so as one o f the differences between different topic crawler is how to calculate the order to crawl URL. Topic crawler can be obtained through the pages of text content to crawl web pages.

3.3 How to enlarge the coverage of theme of web crawler

Topic crawler research trend as theme crawler own studies have advanced. Topic crawler to crawl strategy and algorithm are also constantly improved. The future research direction in the crawler mainly revolve around the following aspects Increase the adaptability of the crawler. The adaptability of the crawler mainly displays in the Internet the organization forms vary considerably between different types of web pages. The web

crawler usually adopts fixed search strategy and it cannot effectively collect various types of web pages. And at the same times topic crawler lacks of adaptability. So how to improve the adaptability of web crawler needs further research.

The initial URL of seed set automatically and the choice of the initial URL seed set of the crawler have an important influence after crawling performance. In general the initial URL seed set of selected aspects often require combination of human and computers to select so as to ensure the efficiency of the crawler. But there are a lot of topic crawler areas every time using artificial and computer combination of time-consuming. How to design the algorithm makes the topic crawler for different areas to automatically generate the corresponding initial URL seed set is a hot research topic in the future. For improving the prediction accuracy to crawl URL theme topic crawler with general crawler of one of the main difference is that can selectively filter has nothing to do with the theme of links and select page orientation on the subject to discover. Therefore, in the treatment of crawling URL topic relevance prediction, if accurately judge to crawl URL, topic relevance and filter irrelevant links, it can greatly improve the efficiency to save time. Topic web crawler application not only in the search engine but also be applied to each topic need to search information. For example, special sites can explore the link between relations and competitive intelligence collection, etc. Topic web crawler mainly has five parts: the data store, download module, pretreatment, web page classification and link analysis. Web page classification is the key to realize subject search. General process of web page classification is selected certain category system and the training data set, then reoccupy taxonomy will be assigned to each category page.

IV. CONCLUSION

This paper mainly introduces the working principle of web crawler and several common web crawler, it also has discussed the topic web crawler. As the main part of search engines, web crawler to crawl the page is a search engine to index in order to achieve quick search. Web crawler technology will be used to deal with an increasing number of network resources and information on the Internet network demand, deal with some of the new technology to develop web pages, crawling some new information. In a word, the web crawler has many problems need us to explore. At the same time, the basic idea of topic web crawler and how to enlarge the coverage of the network problems are worth studying. In addition, the topic crawler to crawl the web process should pay attention to visit the web site of time intervals so as to avoid frequent access to websites bring excessive load.

References

[1]. J. Cho. Crawling the web: Discovery and Maintenance of Large-scale Web Data [D]. L.A.: Stanford University, 2001.
[2]. Jiang Ke. Based on the concept of custom theme crawler system in the field of design and implementation [D]. Xi 'an: xi 'an university of electronic science and technology, 2007.
[3]. Liu Jieqing. Site focused crawler research [D]. Nanchang: jiangxi university of finance and economics, 2006.
[4]. Yu Juan, LiuQiang. Topic web crawler research were reviewed. Computer engineering and science. [J] 5, 2015 (2) : 231-236.
[5]. M Hersocici, M Jacovi, YS Maarek, et al .The shark-search algorithm-An application: Tailored Web site mapping. Proceedings of the 7th International World-Wide Web Conference [C]. Brisbane, Australia:ACM Press,1988.316-325.
[6]. Wang Shuai , Zhou Guo-min, Wang Jian, Reviews of relevance algorithm in focused crawler [J].Computer and Modernization, 2013(4) 27-30.
[7]. S. Chakrabarti, M. van den Berg and B. Dom. Focused Crawling: A New Approach to Topic-Specific Web Resource Discovery [C]. In Proceedings of the 8th International World Wide Web Conference, Toronto, Canada, 1999.
[8]. J. Cho, H. Garcia-Molina. The evolution of the web and implication for an incremental crawler [C].In Proceedings of the 26th International Conference on Very Large Database ,Cairo, Egypt,2000.
[9]. S. Chakrabarti, M. van den Berg and B. Dom. Focused Crawling: A New Approach to Topic-Specific Web Resource Discovery [C].In Proceedings of the 8th International World Wide Web Conference, Toronto Cananda,1999.
[10]. J. Cho, H. Garcia-Molina. The evolution of the web and implications for an incremental crawler [C] .In Proceedings of the 26th International Conference on Very Large Database, Cairo, Egypt, 2000.
[11]. S. Lawrence，C. L. Giles. Accessibility of information on the Web [J]. Nature, 1999.