ISSN 2349-6193
Impact Factor (PIF): 2.243

IJESMR
International Journal OF Engineering Sciences & Management Research

# BUILDING HIGHLY SPECIALISED WEB CRAWLER USING B-TREE SEARCH AND HTML PARSER

**Prateek Raman*, Ravi Kant Gautam, Ravi Yadav, Manish Kumar Sharma**
*1,2,3Graduation Student 4Assistant Professor
1,2,3,4 Department of Computer Science and Engineering, Galgotias College of Engineering and Technology, Greater Noida, India

**DOI**: 10.5281/zenodo.51574

**KEYWORDS:** Best First Search, Priority Strategy of Web Grasping, B-tree Algorithm, Web Revisiting strategy Recommendation System.

## ABSTRACT
Web crawlers are Internet bot that automatically traverse the hyper-link structure of the world wide web in order to locate and retrieve information. This paper describes a web crawling approach based on B-tree search and HTML Parser. As the goal of crawler is to selectively seek out pages that are relevant to given keywords. Rather than collecting and indexing all available web documents to be able to answer all possible queries, a crawler analyze its crawl boundary to hit upon the links that are likely to be most relevant for the crawl, and avoids irrelevant links of the document.

## INTRODUCTION
The World-Wide Web, having over 25 billion pages and associate far growth shows no sign of end. Dynamic content on the web is growing as time-sensitive information such as news,financial data and entertainment become widely dispersed through the web. Search engines are therefore increasingly challenged when trying to uphold current indices using exhaustive crawling. Web crawlers are the heart of search engines, which updates the database as any page gets add or remove from the web. It has become a challenge to traverse all URLs in the web documents and to handle these URLs, because of growing and dynamic nature of the web. A focused crawler is a mediator that targets a particular topic and visits and gathers only relevant web pages.
Web search engines deployed two types of web crawling strategies namely, “breadth” first search and “best” first search. The “best” first search strategy retrieves only those pages which are pertinent to a given topic. Crawler which uses a“best” first search strategy is identified as a “focused crawler. But in this paper we have introduced a new technique to crawl by using “B-tree search” and “HTML Parser”
Our crawler aims at providing a simplest alternative for conquering the issue that instantaneous updation of page which are ranked lowly allied to the given topic at hand. By retrieving those pages which are reachable from the initial seeds, a set of relevant pages is obtained. We find the page which has been updated score with respect to the given topic, from the obtainable set of pages URL Frontier. Set of pages again include this page and its relative pages, from which crawling process will get continue.

## METHODOLOGY USED FOR CRAWLING TECHNIQUE11[2]
The first generation of crawlers [8] on which the majority of the web search engines are based rely heavily on traditional graph algorithms, such as “breadth-first” traversal or “depth- first” traversal, for indexing the web. Document content is paid little need, as the final goal of the crawl is to cover the whole web. Different methodology considering different type of web crawler :-

1. The first and, we might say, the simplest is when the data we are interested can be found on one site at a well defined place or technically record and its value depends on time. For instance this is the logic of storing and presenting market data on some aggregate websites, or showing price information about particular products in a web store or personal information about employees. In these cases, the data structure is static so the crawler has always visit the same page and same record and periodically download its actual value into a database.

2. The second type of crawler is more complex in a sense that it imitates a user who is visiting a list of websites one after the other and downloads data, usually different records from each of these sites which after the completion of the visits is grouped, sorted, and analyzed according to the problem or research

ISSN 2349-6193
Impact Factor (PIF): 2.243

IJESMR
International Journal OF Engineering Sciences & Management Research

design.

3. The third, and most complex, type of search when the browsing sequence is not predetermined neither by its sequence of websites nor the actually numbers of them. This happens typically when we explore a domain of a topic, for instance collect information about graduate programs at universities, looking for a particular service, or try to find out the size and connection structure of a community on the net.

## CORE TECHNOLOGY USED IN WEB CRAWLER[3]

### A. PRIORITY STRATEGY OF WEB PAGE GRASPING

Priority strategy of Web page grasping determines the grasping efficiency. Grasping strategies can be roughly divided into three kinds, i.e. depth-first strategy, breadth-first strategy and best-first strategy. Depth-first strategy could be employed when the amount of information is not huge. However, the rapid development of the Internet and the massive existence of web data will inevitably run into huge data by adopting depth-first algorithm strategy. Therefore, grasping strategies of the search engine will generally be breadth-first strategy and best-first strategy, as well as some of their improved algorithms [2].

### B. DIAMETER OF THE WORLD WIDE WEB

Diameter of the World Wide Web or 'Web Diameter' is defined as 'If d is used to represent a path from Web u to Web v, then the average length of the shortest path formed by all the different pairs of connected pages on the World Wide Web is called Web Diameter. According to this definition and the calculation of large-scale web pages, it can be known that Web Diameter is about 17[3].

The calculation formula of Web Diameter is

$$d=0.35+2.06 \log (N) \quad (1)$$

Study shows that the diameter of China's World Wide Web is 16.26[4], namely if there is a path between any two web pages, click less than 17 times on average, you can reach one web page from another, which is shown in

Figure1.


***Figure 1. Diagram of Diameter of the World Wide Web***

After analyzing the Diameter of the World Wide Web, the following two conclusions are obtained:

(1) Traversing Algorithm has affected the crawler's efficiency to a large extent. The World Wide Web page structure is not that deep as we have imagined, but unexpectedly wider. Therefore, the traversal mode of the crawler generally adopts the breadth-first one. Certainly, there is the reason of the importance of web pages, and this kind of means can help to grasp more important web pages.

(2) The World Wide Web is so complex that a chosen grasping circuit cannot necessarily and invariably guarantee the best. In order to prevent this problem, the diameter of the web needs to be fully considered, and "depth-first strategy" should be adopted to control the grasping depth. In this way, the problem can be perfectly solved [5].

### C. JUDGMENT OF THE WEB IMPORTANCE

While maintaining the priority strategy of web page grasping, please grasp important web pages first to ensure those more important web pages can be arranged with limited resources. Which web pages are more important? How to measure the importance? The measure of importance is decided by the following three aspects, i.e. IB (P), IL(P), ID (P).

1) IB(P)

It is mainly decided by the number and quality of back links. Firstly, the more links (a great many back links) a web page has, the more it is recognized by other pages. Furthermore, there will be more opportunities for it to be visited by net-citizen and its importance is more obvious.

Secondly, the more it is pointed to by more important web pages, the more important it is. The most classic is

ISSN 2349-6193
Impact Factor (PIF): 2.243

IJESMR
International Journal OF Engineering Sciences & Management Research

cheating web pages, which artificially set lots of back-links pointing to their own web pages to increase the importance of web pages. If the quality is not considered, local optimal will appear, rather than problem of global optimal.

2) IL (P)
It is a function of URL string which only investigates the string itself. IL (P) is realized mainly through some models, for example, it attaches more importance to URL containing 'com' or 'home'. It also regards that the URL with fewer slashes is more important.

3) ID(P)
ID (P) represents that in a seed site set; there is a link (breadth-first traverse rules) in every seed site that can arrive at the web page. ID (P) is another important index of the web pages. The closer it is to the seed site, the more opportunities it has to be visited. Therefore, it is more important and the seed site is where the most important web pages are. The farther it is to the seed site, the less important it is.

**D. NON-REPEATED GRASPING STRATEGY**
Massive web page images are other important characteristics of web. According to the 24 million page statistics by Google system, 22% of the web pages are images. The existences of a lot of duplicated web pages are unfavorable to the users' query. It not only wastes the storage space of search engines, but also decreases the system efficiency.

The reasons, on the one hand, are that the collecting program does not clearly record the visited URLs. On the other hand, the domain names and IP addresses have a multiply corresponding relation. The first problem can be solved by making a record of the visited URLs, and making a contrast between the new URLs and the visited ones every time. The second problem is relatively complex, because different URLs may refer to the same IP.

There are four kinds of corresponding relationships between the domain names and IP addresses, namely: one-to-one, one-to-many, many-to-one and many-to-many. One-to-one relationship won't cause repeated collection, but the others are likely to do so.

**1).ALGORITHM BASED ON B-TREE**
Due to the huge amount of web pages, web page grasping requires network bandwidth, machines, time and so on. The repeated grasping of the same web page greatly reduces the efficiency of the system, so the Crawler system should design a strategy to avoid repeated web page grasping to ensure that a web page is grasped only one time in a certain period of time [7].

B-tree is a kind of balanced multi-way search tree. What the file system of operating system uses is the search algorithm of B-tree, which can also be used to design the algorithm matching URL to avoid repeated grasping in the Crawler. B-tree can be empty or multi-way tree. A B-tree of m order must meet the following requirements:
(1) A tree can have m subtrees at most;
(2) If the root node is not the leaf node, at least two subtrees are necessary;
(3) All non-terminal nodes except root have at least two subtrees;
(4) All non-terminal nodes contain the following information data: (n, A0, K1, A1, K2, A2, ..., Kn, An,)subtrees are necessary; Each node includes n pointers pointing to each keyword record. Ki(i=1, ..., n) is keyword and Ki<Ki+1(i=1, ..., n-1); Ai(i=0, ..., n) is the pointer pointing to the root node of subtree and keywords of all nodes of pointer Ai-2 are less than Ki(i=1, ..., n) while keywords of all nodes of the subtree pointed by An are greater than Kn, n($1 \le n \le m$) is the number of keywords(or n+1 is the number of subtrees ). All leaf nodes are in the same layer and carry no information (they can be seen as external nodes or nodes failing to search. Actually,these nodes do not exist and pointers pointing to these nodes are empty).

**E. WEB-PAGE REVISITING STRATEGY**
The popularity of web results from the information web brings. Information is constantly changing, and the web-page information updating is unavoidable. However, the earlier grasped information may be out-of-date or of no use at all. A strategy is thus needed to solve the problem of timeliness of information, and it is called web-page revisiting strategy. Through revisiting, these web-pages can keep pace with the changes of the World Wide

ISSN 2349-6193
Impact Factor (PIF): 2.243

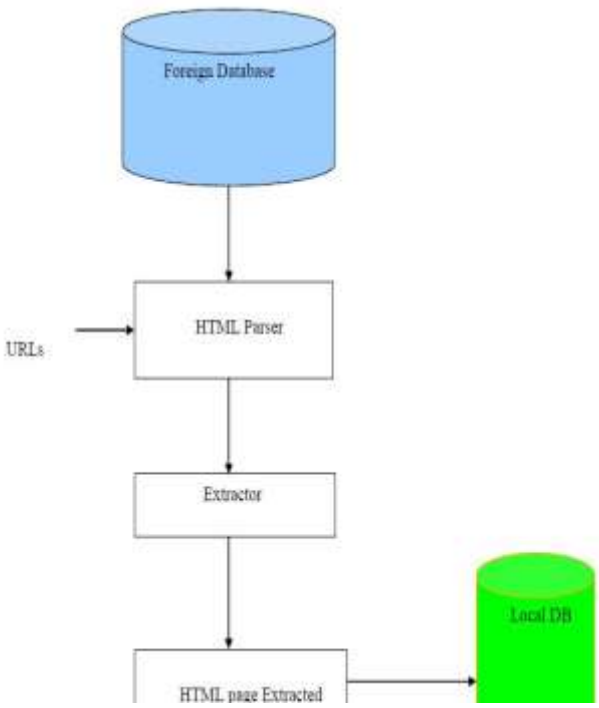
IJESMR
International Journal OF Engineering Sciences & Management Research

Web.
In 2000, Cho and Garcia-Monlina of Stanford University randomly chose 500, 000 web page samples and found that 23% of the web pages were updated on a daily basis while 40% of the web pages with .com as the suffix of their domain names is updated every day. The half-life of web pages is 10 days. In addition, study shows that the process web pages change boils down to model of Poisson process [8].

To describe the model of Poisson process, X(t) is used to represent the number of changes of web pages in the period of (0, t) and the Poisson distribution with λ as its parameter meets the following nature.

**F. DATA EXTRACTION**
In order to extract the data structure from a web-site, we need to identify the relevant fields of information or targets; to achieve that, there are five problems that need to be solved: localizing of HTML pages from various web databases sources; extracting the relevant pieces of data from these pages; distilling the data and improving its structure; ensuring data homogeneity (data mapping); and merging data from various HTML pages into a consolidated schema (data integration problem).

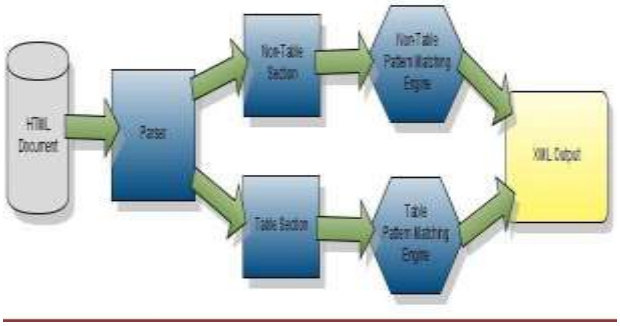
**1).DATA EXTRACTION** from HTML pages Once the relevant HTML pages are returned by various web databases, the system needs to localize the specific region having the block of data that must be extracted without concern for the extraneous information. This can be accomplished by parsing each HTML page returned from different databases. Java has a parser method that parses HTML pages, so this method was used for parsing HTML pages. The foreign database which given in above diagram is SEC database.




**2).PARSING**
The parsing of HTML document to the XML output is as shown in figure below. For the parsing purpose of data we are parse the document from JAVA parser. Then we differentiate both the table and non-table section. Then we pass down the table section using table pattern matching engine.

ISSN 2349-6193
Impact Factor (PIF): 2.243

IJESMR
International Journal OF Engineering Sciences & Management Research




## REFERENCES

1. Maurice de Kunder, (2013) "The Size of the World Wide Web", available from http://www.worldwidewebsize.com/
2. J. Cho, H. Garcia-Molina, and L. Page, "Efficient crawling through URL ordering," in Proceedings of the Seventh World-Wide Web Conference, 1998.
3. M.Najork, J.Wiener, "Breadth-first search crawling yields high-quality pages, " In 10th International World Wide Web Conference, 2001.
4. A.Broker, R.Kumar, F.Maghoul, Tomkins, a.J.Winener, "Graph structure in the web: experiments and models, " presented at Proceedings of the 9th World-Wide Web Conference, Amsterdam, 2000.
5. McCown, F. and Nelson, M. "Agreeing to Disagree: Search Engines and their Public Interfaces". ACM IEEE Joint Conference on Digital Libraries (JCDL 2007). Vancouver, British Columbia, Canada. pp. 309-318. June 17-23, 2007.
6. Hurst M. and Maykov A., (2009) "Social Streams Blog Crawler", In Proceedings of the 2009 IEEE International Conference on Data Engineering
7. Ye S., Lang J., Wu F., (2010) "Crawling Online Social Graphs", In Proceedings of the 2010 Asia Pacific Web Conference.
8. Narayannan Shivakuma, Hector Garcia-Molina, "Finding near-replicas of documents on the web, " Web DB 1998, pp. 204-212.
9. J. Talim, Z. Liu, Ph. Nain, E. G. Coffman. "Controlling the robots of Web search engines, " Proceedings of the 2001 ACM SIGMETRICS international conference on Measurement and modeling of computer systems, Cambridge, Massachusetts, United States, 2001.
10. A.Broker, R.Kumar, F.Maghoul, Tomkins, A.J.Winener, "Graph structure in the web: experiments and models, " presented atProceedings of the 9th World-Wide Web Conference, Amsterdam, 2000
11. S. Chakrabarti, M. van den Berg and B. Dom. "Focused crawling: a new approach to topic specific Web resource discovery", 8 th International WWW Conference, May 1999.
12. Arasu. A, Cho. J, Garcia-Molina. H, "Searching the Web, " ACM Transactions on Internet Technology, pp. 42.
13. Pranit C. Patil, Pramila M. Chawan, Prithiviraj M. Chauhan "Parsing of HTML Document" IJARCET In 2012