IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.918 ⁜ Vol. 11, Issue 11, November 2022

DOI: 10.17148/IJARCCE.2022.111159

Study of Web Crawling Approaches

Dr. Kompal

Govt. College, Panchkula

World Wide Web (WWW) has huge information and latest information is updated so the size of the Web is of order billions of pages. The most significant search engine component is Web crawler. The pages are downloaded continuously and database are stored with pages after indexing. For a crawler, it is not so easy to crawl the complete web and to keep fresh index. To achieve better efficiency, various approaches for Web crawling exist.

1. Web-Crawling Algorithms

The key section of web crawlers is **Scheduler**, which picks up the very next URL for further processing from the list of URL's not visited. Selection of next link largely depends on the crawling algorithm we will use.

There are various approaches to web crawling, each with its own advantages and disadvantages. Here are some common web crawling approaches:

- **Blind Traversing approach**
- **Best – First Heuristic approach**

1.1 Blind Traversing approach

In this, we simply apply the crawling approach procedure on seed URL. It's called blind as no specific criteria is applied while selecting next URL from the frontier. Links from Crawler are selected in the serial order[1]. The algorithm widely used to implement Blind Traversing approach is Breadth First Algorithm. It uses FIFO method to implement the data structure frontier which very basic and simple crawling algorithm.

Since this approach traverses the graphical structure of WWW breadth – wise, Queue data structure is used to implement the Frontier.

Algorithm that comes under Blind Crawling approach is- Breadth First Algorithm.

1.1.1 Breadth First Search Algorithm:

It begins from the root node and looks at all the neighbour nodes at the same level. If the objective is achieved, then it is considered as success and terminates the search else it goes down to next level to look for searching across the neighbour nodes at the same level and keep on doing so till the objective or results are achieved[2]. In case all the nodes are searched and results didn't met the objectives then it's considered as failure.

It uses the frontier as a FIFO queue, crawling links in the order in which they are encountered. The problem with this algorithm is that when the frontier is full, the crawler can add only one link from a crawled page. The Breadth-First crawler is illustrated by following steps

1. Extract Web page
2. Do Parsing
3. Extract various links from the parsed page
4. Add the links to the frontier of FIFO Queue

Breadth-First Algorithm is usually used as a baseline crawler; since it does not use any knowledge about the topic, it acts blindly. That is why, also called, Blind Search Algorithm.

1.2 Best –First Heuristic Approach

To overcome the issues faced in blind traverse approach, a new approach called Best –First Heuristic Approach has been studied by Cho et al.and Hersovici et al. [1998]. In Best –First Heuristic Approach, next link for crawling is taken basis some priority, score or estimates from a given set of links and value of that estimate can be calculated from

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.918 ⁜ Vol. 11, Issue 11, November 2022

DOI: 10.17148/IJARCCE.2022.111159

different mathematical formulas which are predefined[3]. Thus every time the best available link is opened and traversed. In this approach, URL selection process is guided by the lexical similarity between source page and keywords of the URL.

Thus the similarity between a page *p* and the topic keywords is used to estimate the relevance of all the outgoing links of *p*.

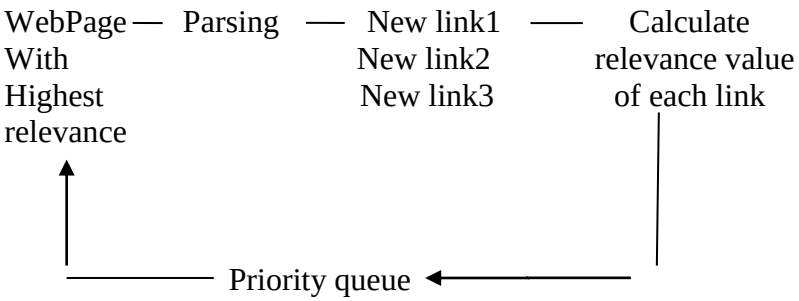

Fig 3.4: Best-First Crawling Approach

Given below Web Crawling Algorithms uses Heuristic Approach:

1.2.1 Naive Best - First Algorithm

The Best First Approach uses relevancy Function rel ( ) which computes the lexical similarity between each given page and chosen keywords & associates the value with matching links in the frontier[4]. After each repetition, the link with highest relevant rel ( ) value is selected from the frontier which means best link available is traversed each time which is not possible in Breadth First Approach. Since links with maximum relevancy value can be selected from the frontier, most of the best first algorithms use **– *Priority Queue*** as data structure.

1.3 Page Rank Algorithm

Page Rank Algorithm decides the significance of the web pages by calculating citations or backlinks to a given page . The page rank of a given page is calculated as
PR(A) = (1-d) + d (PR(T1)/C(T1) + ... +
PR(Tn)/C(Tn))
PR(A) : Page Rank of a Website,
d : damping factor
T1,….Tn : links
The Page Rank of a web page is therefore calculated as a summation of the page ranks of the pages linking to it (its incoming links), divided by the number of links on each of the respective pages (its outgoing links).

1.4 Fish – Search Algorithm

This algorithm treats Internet as a directed graph, hyperlink as edge and webpage as node, so that the search could be abstracted as process of traversing graph. For every node, we decided whether its relative or not i.e. 1 for relevant and 0 for irrelevant. This algorithm keeps a list which maintains URLs of the pages to be searched. Each URL has different priority; each URL will be assigned priority and URLs with higher priority will be shown first in the list and will be searched before than others.

1.5 Shark Search Algorithm

In Shark Search Algorithm, one direct improvement is that it returns a “fuzzy” Score i.e. score will be 0 and 1 (0 for zero resemblance, 1 for exact “conceptual” match) instead of binary evaluation of document relevance.

This algorithm is an enhanced version of Fish Search algorithm. The child inherits the discounted value of ancestors and Shark Search also keeps in consideration the Anchor text of a web page while assigning it any relevancy value.

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.918 ⁜ Vol. 11, Issue 11, November 2022

DOI: 10.17148/IJARCCE.2022.111159

CONCLUSION

The choice of web crawling approach depends on the specific goals and constraints of the crawling project. Successful web crawling requires a combination of technical expertise, domain knowledge, and careful planning. As computing resources are limited and time constraints, the various types of crawlers came into picture.

REFERENCES

[1]. Junghoo Cho and Hector Garcia-Molina "Effective Page Refresh Policies for Web Crawlers "ACM Transactions on Database Systems, 2003.

[2]. Sandeep Sharma and RavinderKumar, "Web-Crawlers and Recent Crawling Approaches," International Conference on Challenges and Development (IT-ICCDIT-2008), PCTE, Ludhiana (Punjab), May 30th, 2008.

[3]. Dhiraj Khurana, Satish Kumar, "Web Crawler: A Review", IJCSMS International Journal of Computer Science & Management Studies, Vol. 12, Issue 01, January 2012.

[4]. Yugandhara Patil, Sonal Patil, Janar 2016 "Review of Web Crawlers with Specification and Workin"g Vol. 5, Issue 1, January 2016

