Literature review on Web Crawling

Sarvesha Chodankar
School of Computer Science
Dr. Vishwanath Karad MITWPU
Pune, Maharashtra

Amanda Michael
School of Computer Science
Dr. Vishwanath Karad MITWPU
Pune, Maharashtra

Siddharth Walke
School of Computer Science
Dr. Vishwanath Karad MITWPU
Pune, Maharashtra

Dr. C.H.Patil
School of Computer Science
Dr. Vishwanath Karad MITWPU
Pune, Maharashtra

***Abstract* - This work provides a literature review on Web Crawling. A web crawler is a software program which browses the World Wide Web, creates a list of web pages, and indexes it so that the required information can easily be retrievable. This process is called Web Crawling. This paper includes brief details about web crawling.**

Keywords- crawler, Wandex, MOMspider, PageRank, Deep web, focused crawling, path-ascending crawling.

I. INTRODUCTION

Web Crawlers are also called spiders or bots. Web crawling basically maps the internet as a graph where the nodes are the web pages and the edges are links showing how they are related to each other. Hence they are used by search engines to not only index web pages and to keep database up to date but also tests websites' vulnerability.

Web Crawlers can crawl through only public pages on websites and not the private pages which are referred to as "dark web".[1]

The search engines highly rely on the web

crawlers because the amount of information on internet is increasing day by day as the new pages are added or updated, to learn about this updates or new data the web crawler constantly crawls, keeps on updating the search engine database and provides better results to search engines.[3]

II. HISTORY

In 1993, the first crawler named 'World Wide Web Wanderer' was introduced .[2]In the same year more 3 crawlers were introduced named 'Jump Station', 'World Wide Web Worm' and 'RBSE spider'.[3] Initially these crawlers were used to compute the size of the web ,later they were used to retrieve URLs from the first websearch engine named 'Wandex'.[2] In 1994,'WebCrawler' and 'MOMspider' were introduced. In just one year the number of pages raised from 110,000 to 2 million. Later a few more crawlers such as Lycos, Infoseek, Excite, AltVista and HotBot became available. In 1998 web crawler 'Google' was introduced which reduced the disk access time through compression and indexing techniques and also used an algorithm named 'PageRank' which calculated the probability of a user visiting a page.

III. WORKING OF WEB CRAWLER

The very first step of web crawling is to begin with seed URLs(a set of URLs). The Crawler moves from one link to another, extracting new links available in the downloaded pages. Then the retrieved pages are stored in the database in a well indexed manner so that they can easily be retrieved in future. It is checked if URLs's related documents are downloaded or not and the process of downloading documents is repeated till all URLs are downloaded.[4].

Before starting with the crawling process the crawler is supposed to read a file called "robots.txt" which instructs crawler which parts of the website to ignore. In a way robots.txt file provides a way to control what crawler can see on a website. Crawler eventually deletes a page from index when it can't find it however some crawlers check for pages twice in such cases.[2]

The working of web crawler can be defined in steps as follows:

Selecting a starting seed URL or URLS
Fetching the web page corresponding to selected URL
Parsing the same web page to find new URL links
Adding new URLs to database
Go to step 2 and repeat the process until no more URLs are left.[4]

CHARACTERISTICS OF WEB CRAWLER

Distinguishing and removing duplicate data obtainable from various websites.[6]

Capable of executing on various machines simultaneously.[6] Compatibility with static as well as dynamic pages.[6] Continuous crawling [1]

IV. TYPES OF WEB CRAWLER

A. General-Purpose Web Crawler

The crawlers collect and fetches the entire contents of web and store it in a centralized location so they can be indexed in advance.[2]

B. Incremental Web Crawler

The crawlers keep on updating the set of pages and store their recent copy. The crawler keeps track of available pages, checks if they have changed from last time and updates accordingly.[6]

C. *Deep web crawler*

There are more than 90% of data on the web that are not even accessible by the crawlers, such type of data is known as deep/dark web. To deal with such data deep web crawlers are designed.[6]

D. *Distributed Crawler*

The crawling process is divided into multiple processes due to an increase in data on the web. The distributed web crawler multiple crawlers download web pages in parallel and resulted pages are sent to a central indexer where the links are extracted and sent to the crawlers via the URL server.[6]

E. *Parallel web crawler*

Sometimes it is difficult to retrieve a web page using a single process therefore to get the required page, search engines run multiple processes parallelly. This is done by parallel web crawler.[2]

F. *Adaptive Crawler*

The adaptive crawler uses previous cycles of crawling to check which pages should be updated. This crawler is called as an incremental type of crawler.[2]

G. *Breadth First crawler*

The crawling process is done in breadth-first fashion.[2]

V. ALGORITHM FOR WEB CRAWLING

Different Algorithms to handle web crawling procedure efficiently:

A. *Focused Crawling:*

This algorithm downloads pages that are similar to each other, also known as Topical or Focused crawler. The context of the page is an important aspect for this algorithm. The main task of this algorithm is to predict the similarities between pages before downloading them. The focused crawling algorithm's performance depends mostly on the richness of the links in a specific topic being searched.[5]

B. *Path - Ascending Crawling:*

Path-ascending crawler helps in finding isolated pages which the regular crawler would have not found. [5] The path ascending crawler ascends to every path in each URL and download all possible documents from website

CONCLUSION

The data on the internet is increasing rapidly and will continue to increase in future. Hence a good crawling algorithm is important which will function rapidly, at the same time will give better results. They are often considered as the basic component of web services. There are already a number of crawling algorithms in the market used by various search engines. Choosing the right strategies and architecture implementation is a main task in building a web crawler.

REFERENCES

[1] Mini Singh Ahuja, Dr Jatinder Singh Bal, Varnica." Web Crawler: Extracting the web data"
[2] S.Amudha. "Web Crawler for mining web data"
[3] Seyed M. Mirtaheri, Mustafe Wmre Dincturk,Salman Hooshman, Gregor V. Bochmann, Guy-Vincent Jourdan, "A Brief History of Web Crawlers"
[4] Md. Abu Kausar, V.S. Dhaka, Sanjeev Kumar Singh. "Web Crawler:A Review"
[5] Yaduvir Singh "Review paper on Web Crawler"
[6] Vandana Shrivastava "A Methodical Study of Web Crawler.