International Journal of Electronics Communication and Computer Engineering
Volume 4, Issue 3, ISSN (Online): 2249–071X, ISSN (Print): 2278–4209

Study of Web Crawler

Siddharth Bhandari
Email: Siddharth.bhandari23@gmail.com

Simal Sethi
Email: Simal.sethi@gmail.com

Suyog Gune
Email: Suyog.gune07@gmail.com

Abstract **- A web crawler is a relatively simple automated program, or script that methodically scans or crawls through Internet pages to create an index of the data it's looking for; these programs are usually made to be used only once, but they can be programmed for long-term usage as well. There are several uses for the program, perhaps the most popular being search engines using it to provide webs surfers with relevant websites. Other users include linguists and market researchers, or anyone trying to search information from the Internet in an organized manner. Alternative names for a web crawler include web spider, web robot, crawler and automatic indexer. Crawler programs can be purchased on the Internet, or from many companies that sell computer software, and the programs can be downloaded to most computers.**

Keywords **- Crawlers, Frontier, Seed URL, Web Search Engines.**

I. INTRODUCTION

A web-crawler is a program/software or an automated script which browses the World Wide Web in a methodical, automated manner. The structure of the World Wide Web is a graphical structure, *i.e*. the links given in a page can be used to open other web pages. Actually Internet is a directed graph, webpage as node and hyperlink as edge, so the search operation could be abstracted as a process of traversing directed graph. By following the linked structure of the Web, we can traverse a number of new web-pages starting from a Starting webpage. Web crawlers are the programs or software that uses the graphical structure of the Web to move from page to page. Such programs are also called wanderers, robots, spiders, and worms. Web crawlers are designed to retrieve Web pages and add them or their representations to local repository/databases. Web crawlers are mainly used to create a copy of all the visited pages for later processing by a search engine that will index the downloaded pages that will help in fast searches. Web search engines work by storing information about many web pages, which they retrieve from the WWW itself. These pages are retrieved by a Web crawler (sometimes also known as a spider) which is an automated Web browser that follows every link it sees. Web *crawlers* are programs that exploit the graph structure of the web to move from page to page. It may be observed that `crawlers' itself doesn't indicate speed of these programs, as they can be considerably fast working programs. Web crawlers are software systems that use the text and links on web pages to create search indexes of the pages, using the HTML links to follow or crawl the connections between pages.

II. A SURVEY OF WEB CRAWLERS

Web crawlers are almost as old as the web itself. The first crawler, Matthew Gray's Wanderer, was written in the spring of 1993, roughly coinciding with the first release of NCSA Mosaic. Several papers about web crawling were presented at the first two World Wide Web conferences. However, at the time, the web was three to four orders of magnitude smaller than it is today, so those systems did not address the scaling problems inherent in a crawl of today's web.

The original Google crawler (developed at Stanford) consisted of five functional components running in different processes. A *URL* server process read URLs out of a file and forwarded them to multiple crawler processes. Each crawler process ran on a different machine, was single-threaded, and used asynchronous I/O to fetch data from up to 300 web servers in parallel. The crawlers transmitted downloaded pages to a single Store Server process, which compressed the pages and stored them to disk. The pages were then read back from disk by an indexer process, which extracted links from HTML pages and saved them to a different disk file. A *URL* resolver process read the link file, derelativized the URLs contained therein, and saved the absolute URLs to the disk file that was read by the URL server. Typically, three to four crawler machines were used, so the entire system required between four and eight machines.

Research on web crawling continues at Stanford even after Google has been transformed into a commercial effort. The Stanford Web Base project has implemented a high- performance distributed crawler, capable of downloading 50 to 100 documents per second .Cho and others have also developed models of document update frequencies to inform the download schedule of incremental crawlers.

The Internet Archive also used multiple machines to crawl the web. Each crawler process was assigned up to 64 sites to crawl, and no site was assigned to more than one crawler. Each single-threaded crawler process read a list of seed URLs for its assigned sites from disk into per-site queues, and then used asynchronous I/O to fetch pages from these queues in parallel. Once a page was downloaded, the crawler extracted the links contained in it. If a link referred to the site of the page it was contained in, it was added to the appropriate site queue; otherwise it was logged to disk. Periodically, a batch process merged these logged "cross-site" URLs into the site-specific seed sets, filtering out duplicates in the process.

International Journal of Electronics Communication and Computer Engineering
Volume 4, Issue 3, ISSN (Online): 2249–071X, ISSN (Print): 2278–4209

III. WEB CRAWLER ARCHITECTURE

This section provides an overview of how the whole system of a search engine works. The major functions of the search engine crawling, indexing and searching are also covered in detail in the later sections. Before a search engine can tell you where a file or document is, it must be found. To find information on the hundreds of millions of Web pages that exist, a typical search engine employs special software robots, called spiders, to build lists of the words found on Websites. When a spider is building its lists, the process is called Web crawling. A WebCrawler is a program, which automatically traverses the web by downloading documents and following links from page to page.

They are mainly used by web search engines to gather data for indexing. Other possible applications include page validation, structural analysis and visualization; update notification, mirroring and personal web assistants/agents etc. Web crawlers are also known as spiders, robots, worms etc .Crawlers are automated programs that follow the links found on the web pages. There is a URL Server that sends lists of URLs to be fetched to the crawlers. The WebPages that are fetched are then sent to the store server. The store server then compresses and stores the web pages into a repository. Every web page has an associated ID number called a doc ID, which is assigned whenever a new URL is parsed out of a web page. The indexer and the sorter perform the indexing function. The indexer performs a number of functions. It reads the repository, uncompressed the documents, and parses them. Each document is converted into a set of word occurrences called hits. The hits record the word, position in document, an approximation of font size, and capitalization. The indexer distributes these hits into a set of "barrels", creating a partially sorted forward index. The indexer performs another important function. It parser out all the links in every web page and stores important information about them in an anchors file. This file contains enough information to determine where each link points from and to, and the text of the link. The URL Resolver reads the anchors file and converts relative URLs into absolute URLs and in turn into doc IDs. It puts the anchor text into the forward index, associated with the doc ID that the anchor points to. It also generates a database of links, which are pairs of doc IDs. The links database is used to compute Page Ranks for all the documents. The sorter takes the barrels, which are sorted by doc ID and resorts them by word ID to generate the inverted index. This is done in place so that little temporary space is needed for this operation. The sorter also produces a list of word IDs and offsets into the inverted index. A program called Dump Lexicon takes this list together with the lexicon produced by the indexer and generates a new lexicon to be used by the searcher. A lexicon lists all the terms occurring in the index along with some term-level statistics (e.g., total number of documents in which a term occurs) that are used by the ranking algorithms. The searcher is run by a web server and uses the lexicon built by Dump Lexicon together with the inverted index and the Page Ranks to answer queries.

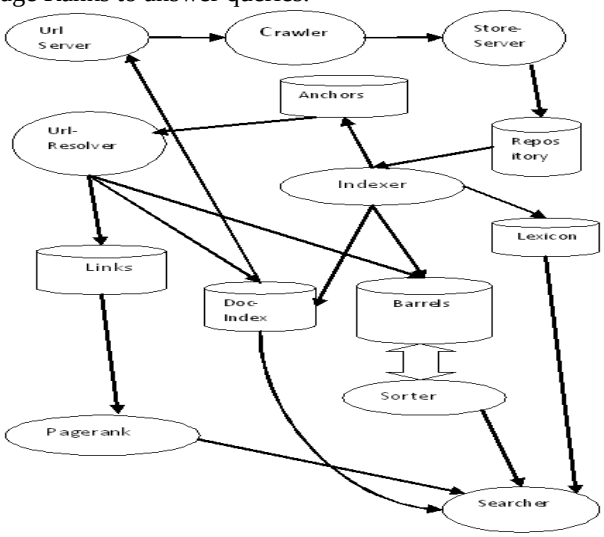

Fig. Web Crawler Architecture

IV. CONCLUSION

Internet is one of the easiest sources available in present days for searching and accessing any sort of data from the entire world. The structure of the World Wide Web is a graphical structure, and the links given in a page can be used to open other web pages. In this Paper, we have used the graphical structure to process certain traversing algorithms used in the search engines by the Crawlers. Each webpage can be considered as node and hyperlink as edge, so the search operation could be abstracted as a process of traversing directed graph. By following the linked structure of the Web, we can traverse a number of new web-pages starting from a Starting webpage. Web crawlers are the programs or software that uses the graphical structure of the Web to move from page to page.

REFERENCES

[1] Baldi, Pierre. Modeling the Internet and the Web: Probabilistic Methods and Algorithms, 2003.
[2] Arvind Arasu, Junghoo Cho, Andreas Paepcke: "Searching the Web", Computer Science Department, Stanford University.
[3] David Hawking, "Web Search Engines: Part 1", June 2006
[4] Gautam Pant, Padmini Srinivasan, and Filippo Menczer: "Crawling the Web" available at-<< http://dollar.biz.uiowa.edu/~pant/Papers/crawling.pdf>>
[5] Marc Najork, Allan Heydon SRC Research Report 173, "High-Performance Web Crawling", published by COMPAQ systems research center on September 26, 2001.
[6] Sergey Brin and Lawrence Page, "The anatomy of a large-scale hyper textual Web search engine", In *Proceedings of the Seventh International World Wide Web Conference*, pages 107–117, April 1998.
[7] Sandeep Sharma and Ravinder Kumar, "Web-Crawlers and Recent Crawling Approaches", in International Conference on Challenges and Development on IT - ICCDIT-2008 held in PCTE, Ludhiana (Punjab) on May 30th, 2008
[8] Junghoo Cho, Hector Garcia-Molina: "Parallel Crawlers", 7–11 May 2002, Honolulu, Hawaii, USA.

[9] Marc Najork, Janet L. Wiener," Breadth-first search crawling yields high-quality pages", WWW10 proceedings in May 2-5, 2001, Hong Kong.

[10] Filippo Menczer, Gautam Pant and Padmini Srinivasan, "Topical Web Crawlers: Evaluating Adaptive Algorithms" ACM Transactions on Internet Technology, Vol. 4,

[11] M. Hersovici, M. Jacovi, Y. Maarek, D. Pelleg, M. Shtalhaim and S. Ur. "The Shark-Search Algorithm – An Application: Tailored Web Site Mapping", In *Proceedings of the Seventh International World Wide Web Conference*, Brisbane, Australia, April 1998. Available at- << http://www7.scu.edu.au/1849/com1849.htm >>

[12] A. Z. Broder, S. R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. L. Wiener, "Graph structure in the web" in Proc. Of WWW Conf., 2000 Availableat<http://www.cis.upenn.edu/~mkearns/teaching/NetworkedLife/broder.pdf>

[13] Dr. P.M.E. De Bra, Drs. R.D.J. Post, "Searching for arbitrary information in the WWW: the fish-search for Mosaic."Availableat<<http://archive.ncsa.uiuc.edu/SDG/IT94/Proceedings/Searching/debra/article.html >>

[14] Blaz Novak, "A survey of focused web crawling algorithms", Department of Knowledge Technologies, Jozef Stefan Institute, Ljubljana, Slovenia

[15] Edleno S. de Moura, Daniel R. Fernandes, Altigran S. Silva, "Improving Web Search Efficiency via a Locality Based Static Pruning Method", *WWW 2005*, May 10-14 2005, Chiba, Japan.

AUTHOR'S PROFILE

Siddharth Bhandari
Pursuing B.E. (C.S). from Sanghvi Institute of Management & Science.

Simal Sethi
Pursuing B.E(C.S). from Sanghvi Institute of Management & Science.

Suyog Gune
Pursuing B.E.(C.S.) from Sanghvi Institute of Management & Science.