I J C T A, 9(34) 2016, pp. 229-234

Extracting Information Using Effective Crawler Through Deep Web Interfaces

J. Jayapradha*, D. Vathana and D.Vanusha*****

ABSTRACT

The World Wide Web is a vast collection of billions of web forms containing thousands of gigabytes of data arranged in many servers using HTML. A Web Crawler is an internet bot which surf the World Wide Web, for the purpose of Web spidering. In the early years, even the largest crawlers fail in making the absolute index as the number of web pages is tremendously large. Nowadays, modern search engines give instant results with irrelevant links. Here, the aim is to create an efficient crawler, which eliminates the irrelevant links, only gives the relevant links for the given data. In this paper, we present the architecture to eliminate the irrelevant data. Prioritization of links is done according to the ranking system through site prioritization methods. Crawlers are used to create the copy of all web pages visited and processed by the search engine, which further will be indexed the downloaded pages, which helps in quick searches. As a result, the prioritization of relevant links is achieved by eliminating the irrelevant links through Site prioritization method and also provides the graphical representation of the relevant data through the ranking system for the higher priority records to higher view count by the user.

Keywords: World Wide Web, search engine, web crawler, database, site prioritization.

INTRODUCTION

The World Wide Web is an enormous collection of billions of web forms containing thousands of gigabytes of data arranged in many servers using HTML [2]. It is impossible to find a particular data in huge cloud available. The number of irrelevant data compared to relevant data is comparatively high, and another problem in searching data is less accurate results., this may cause by keystrokes typed by a user has many contexts, and users do not need most of them. For this reason of the huge number of irrelevant web pages resulting less accurate results, web crawlers are used.

A web crawler is a system that goes around the web gathering and putting away [2] information in a database for further investigation and a course of action. The procedure of web crawling includes gathering pages from the web and masterminding them in a manner that the web index can recover them efficiently. The critical objective is to do so efficiently and quickly without much interference with the functioning of the remote server.

A Web crawler begins with a URL Seeds, which are URL or sets of URL. The crawler visits the URL to which it looks like the hyperlink to other web pages; it adds the URL to an existent file [3] [4], and this methodology of visiting the URL depends on rule sets for the crawler. The data collected is sent to a database for further analysis.

There are two basic kinds of crawlers, generic-crawler which is to find all forms of data without any specifications; here the data related to the keyword are specified ambiguously. Focused Crawler method, web crawler searches the data specifically on the certain topic [5]. They manage the hyperlink exploration process; focused crawlers predict the infested sites may be useful even before downloading it.

* Asst. Prof (O.G), Dept of Computer Science, SRM University, India, *E-mails: jayapradha.j@ktr.srmuniv.ac.in; vathana.d@ktr.srmuniv.ac.in; vanusha.d@ktr.srmuniv.ac.in*

A focused crawler downloads a page only when the page is similar to the result; they are also called tropical crawlers.

2. BACKGROUND OF CRAWLER

A web crawler is a tool which is an integral part of search engines, which surfs through the web finding data and store them into databases for further enhancements and analysis of data. They are also called web spider, ant, robot, and bot. Crawler which is a robot, which helps to find the file or URL or any sets of data.

The background of crawler goes from generic-crawler which is to find all forms of data without any specifications; here the data related to the keyword are specified ambiguously. The next updated version of the web - crawler is focused crawler, which focused on particular sets of forms.

In focused crawler method, web crawler searches the data specifically on the certain topic [6]. They manage the hyperlink exploration process; focused crawlers predict the infested sites may be useful even before downloading it. A focused crawler downloads a page only when the page is similar to result, and they are also called tropical crawlers. The performance of focused web crawler is based on the virtual search on a particular topic. It crawls pages from a general search engine.

Some of these crawlers are general-purpose crawlers which can be used in daily life are Bingbot, FAST Crawler, Googlebot, PolyBot, Swiftbot and World Wide Web Worm.

2.1. Generic Web Crawler

The generic web crawler is a typical web crawler. It gathers all the irrelevant links to the data that is specified. Generic doesn’t prioritize links according to the particular subject.

2.2. Focused Web Crawler

Focused Crawler is the Web crawler that gathers web pages on a particular data [7]. It gathers documents which are meticulous and important to the given topic. It is otherwise called a Topic Crawler on account of its method for working. It decides how far the given page applies to the particular point and how to continue forward. The benefits of focused web crawler are that it is economically feasible regarding network and hardware resources; it can decrease the amount of network traffic and downloads. The inquiry representation of focused web crawler is also massive.

2.3. Incremental Crawler

An Incremental Crawler, to refresh its collection, periodically replaces the old documents with the newly downloaded documents. On the contrary, an incremental crawler incrementally updates the existing collection[6] of pages by visiting them regularly; based upon the estimate as to how often pages change. It also exchanges less important pages of new and more relevant pages. It determines the issue of the freshness of the pages. The advantage of the incremental crawler is that just the client gets the valuable information so the data enrichment is accomplished and the network bandwidth is retained.

2.4. Distributed Crawler

Distributed web crawling is a distributed computing technique. Numerous crawlers are working to spread in the process of web crawling, to have the most coverage of the network. A central server is geographically[5] distributed which manages the communication and synchronization the nodes. It uses Page Rank algorithm for its increased efficiency and quality search. The distributed web crawler is robust.

2.5. Parallel Crawler

Multiple crawlers frequently keep running in parallel, referred as Parallel crawlers. A crawler consists of various crawling Processes called as a C - process which can run on a network of workstations. The Parallel crawlers rely on upon Page freshness and Page Selection. A Parallel crawler can be on the local network or distributed, at geographically distant locations. Parallelization of the crawling system is very much essential for downloading documents in a reasonable amount of time.

3. DESIGN OF DATABASE

We break up our work into two parts: Front-End Development and Back-End Development. A back-end developer is the responsible for the server-side web application logic and integration of the front-end. Back-end developers usually carry out the web services and APIs used by front-end developers. The back-end tool is a part of the application that is never noticeable to the user and built with the server-side languages [8]. In simpler words; front-end code interacts with a user in real time while the back-end system communicates with the server to return user results. Anything displayed on the website performs because of a query on the server returned data to the front-end.

The operation of the back-end code is a bit more complicated. The developer creates an application (using server-side code, like PHP, Html, CSS, Java, etc.) which will connect to a database (using MySQL, SQL, Access, etc.) to know, save or change data and return it back to the user in the form of front-end code. This complicated structure helps us to look for things, shop, interact socially, and more in the modern world of the internet. Back-end Code is responsible for any operation. The back-end developer usually takes the ready front-end code and implements it into the application where everything is displayed dynamically using the data stored in a database. This developer works closely with proposed functionality specifications that the end user wants to achieve. Back end development frequently contains the API (Application Programming Interface) code. A couple of database operations and possible a few uniform front end pages with a superior utilization of wide end to end functionality, which is the perfect way to understand back-end usefulness.

3.1. Sqlyog community

Sqlyog community is freely trusted, secured software for a database. Also, SQL works on many front-end platforms. Also, it is the most powerful MySQL administration tool for the database. This tool makes related to maintaining relational databases efficiently, most effectual and fast. It is a GUI administration tool for MYSQL developed on databases. Also, it has a secured login for accessing.

We can bind our MySQL services to local host, no need to open the service to listen on all IP addresses. We can access the old local host itself, there is no need of adding new users. We can even use our root user to do the same management tasks. Also, in this one importing and updating files quickly, the other users may not have permission to access and create more users, the new user will not other access to the server. The new user needs permission to access. Actually, in this MySQL, many users can use this one, but for PHP users each and every user has the separate account. MySQL is a good GUI. This application will be the user interface to our database. The use of Sqlyog is data will be building and manipulating. Sqlyog is useful to view the data on a remote web server that we are maintaining. Also in this one drop down list option is there to use to choose the various variable types.

4. SYSTEM ARCHITECTURE

Working of Architecture

The Architecture of efficient crawler from figure 1 consists of 5 components, namely Crawler, Link Fetcher, Link Frontier, Page Fetcher and Database

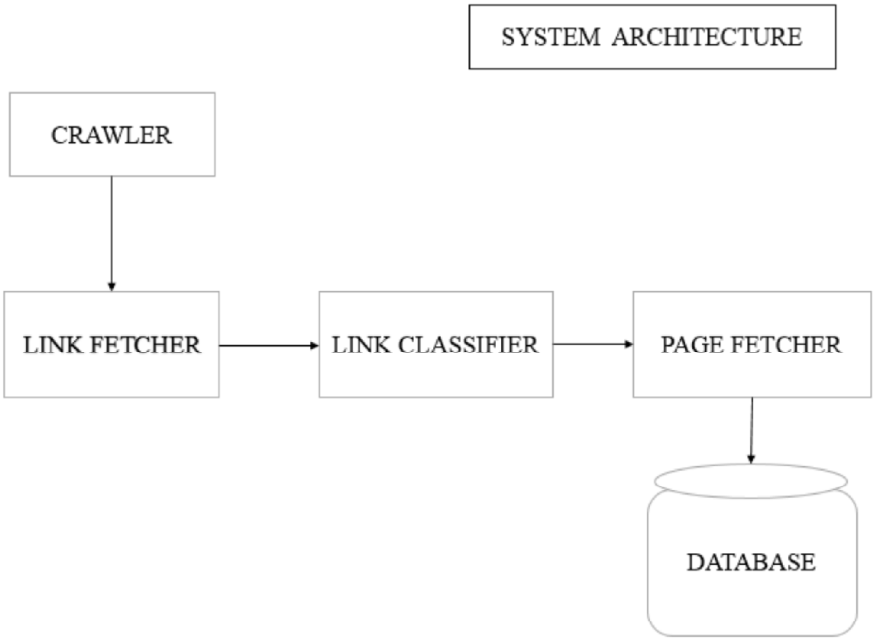

Figure 1: System Architecture

When a user types the keyword for the data to be found, this follows five steps to eliminate irrelevant data and provide relevant data for particular key search. First, a keystroke is given to the crawler to find the appropriate web forms. By the reverse searching method, web crawler crawls all the data. Link fetcher fetches the links provided by the crawler [12] Link classifier takes all the links from the link fetcher and prioritizes virtual links, by process of site prioritization. Page fetcher fetches the page from the data provided from link classifier. The database stores all the information in the backend. The logs of the user's view count and stores the record of relevant sites.

4.1. Crawler

A crawler is a tool which is an integral part of search engines, which surfs through the web finding data and store them into databases for further enhancements and analysis of data [13]. They are also called web spider, ant, robot, and bot. Crawlers are used to create the copy of all web pages visited and processed by the search engine, which further will be indexed the downloaded pages, which helps in quick searches. In this architecture crawler takes inputs of keystrokes, these keystrokes must consist of alphabets, numbers and special characters only. By keystrokes, it finds the virtual search on the internet.

4.2. Link Fetcher

Link fetcher module fetches the links provided by the data given by Bot. Link fetcher fetches all the links which are relevant and unrelated for particular keystroke and stores it. According to a meticulous topic, Link fetcher gains all the links.

4.3. Link Classifier

Link classifier takes the entire links from link fetcher, and it prioritizes the relative links ignoring the unrelated links. By process of site prioritization, link classifier prioritizes the relative links among all other links.

4.4. Page Fetcher

It fetches the page from the data provided from link classifier. The page fetcher allows a user to view the link.

4.5. Database

The database stores all the records and the users visited logs. For reference purpose the keystrokes are also stored and after prioritization of sites the record of a relevant site are saved in a database, and sent to link classifier.

5. LITERATURE SURVEY

World Wide Web, its size, its structure and storage of data [2] where data is a dynamic and problems in retrieving this data is due to traffic patterns.

The literature review [3] describes and explains how the Web crawler works, how does it crawls the data and retrieves it, its components present in web crawler and also modules used.

As per the survey, Generic Framework is used for Specifying the User Interest and for Adaptive Crawling [4] Strategies helps in finding results of searching.

In the survey of web crawling to analyze page ranking algorithm [6], several algorithms like site ranking algorithm, Breadth-First Search Algorithm and Depth First Search Algorithm was discussed.

Two stage crawlers [1] were introduced for the efficient crawling of deep web forms by implementing the site locating and in site exploration methods. Reverse searching algorithm and site prioritization algorithm were also used for getting the accurate and prioritization of data.

6. GRAPHICAL REPRESENTATION

In figure 2 Graphical representation of the result represents, according to priority of links, with the higher priority records to higher view count by the user.

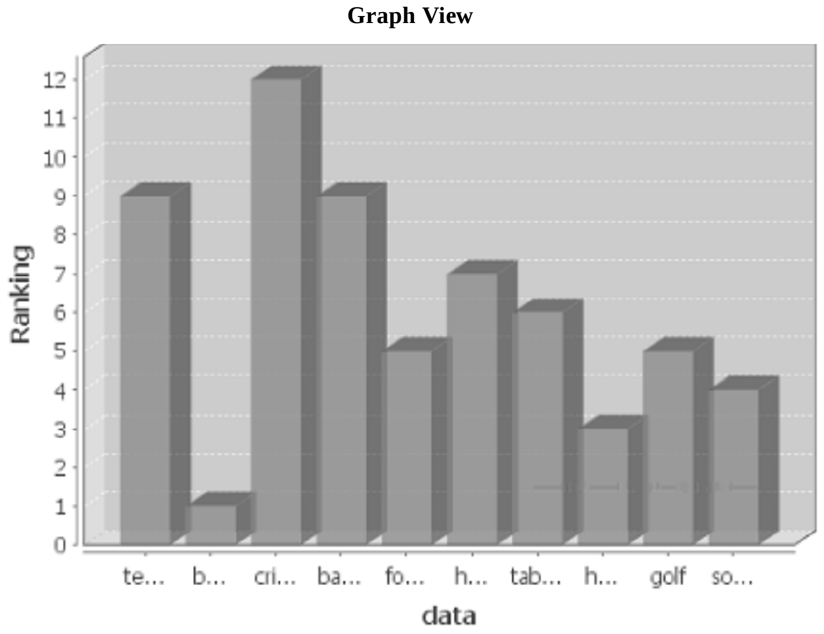

Figure 2: Graph View

7. RESULT AND FUTURE WORK

In this paper, we propose an efficient crawler, which eliminates the irrelevant links provided, gives only relevant links for the given data. We also provide the graphical representation of the result according to priority of the links. In a world wide web, there are thousands of gigabytes of data stored in the cloud, finding particular reliable information is hard. There is always a chance of data having many domains; this will result in many non-relevant links as a result. In the future, we plan to enhance the accuracy of crawler by applying pre-query and post-query for web forms.

REFERENCES

[1] Smart Crawler: A Two-stage Crawler for Efficiently Harvesting Deep-Web Interfaces Feng Zhao, Jingyu Zhou, Chang Nie, Heqing Huang, Hai Jin, IEEE Transactions on Services Computing Volume: PP Year: 2015.

[2] Abha Joshi , Avani Jadeja Improving Algorithm for Calculation of Page Rank, The International Journal Of Science & Technoledge (ISSN 2321 – 919X) May, 2015.

[3] Md. Abu Kausar, V. S. Dhaka and Sanjeev Kumar Singh, Design of Web Crawler for the Client - Server Technology, Indian Journal of Science and Technology, *Vol 8(36), DOI: 10.17485/ijst/2015/v8i36/79858, December 2015.*

[4] Paras Nath Gupta, Pawan Singh, Pankaj P Singh, Punit Kr Singh, Deepak Sinha , A Novel Architecture of Ontology based Semantic Search Engine , International Journal of Science and Technology Volume 1 No. 12, December, 2012.

[5] iRobot: An Intelligent Crawler for Web Forums Rui Cai, Jiang-Ming Yang, Wei Lai, Yida Wang, and Lei Zhang , Microsoft Research, Asia. April 21-25, 2008 · Beijing, China.

[6] Survey of Web Crawling Algorithms Rahul Kumar, Anurag Jain and Chetan Agrawal, Advances in Vision Computing: An International Journal (AVC) Vol.1, No.2/3, September 2014.

[7] Soumen Chakrabarti, Martin Van den Berg, and Byron Dom. Focused crawling: a new approach to topic-specific web resource discovery. Computer Networks, 31(11):1623–1640, 1999.

[8] Yeye He, Dong Xin, Venkatesh Ganti, Sriram Rajaraman, and Nirav Shah. Crawling deep web entity pages. In Proceedings of the sixth ACM international conference on Web search and data mining, pages 355–364. ACM, 2013.

[9] Jayant Madhavan, David Ko, £ucja Kot, Vignesh Ganapathy, Alex Rasmussen, and Alon Halevy. Google’s deep web crawl. Proceedings of the VLDB Endowment, 1(2):1241–1252, 2008.

[10] Shestakov Denis. On building a search interface discovery system. In Proceedings of the 2nd international conference on Resource discovery, pages 81–93, Lyon France, 2010. Springer.

[11] World Wide Web network traffic patterns, Compcon ’95.’Technologies for the Information Superhighway’, Digest of Papers, Author: J. Sedayao, 5-9 March 1995 , ISSN 1063-6390.

[12] Web Crawling By Christopher Olston and Marc Najork , Foundations and Trend in Information Retrieval Vol. 4, No. 3 (2010) 175–246 2010 C. Olston and M. Najork DOI: 10.1561/1500000017.

[13] Focused Web Crawling: A Generic Framework for Specifying the User Interest and for Adaptive Crawling Strategies(2001), Martin Ester, Matthias Groß, Hans-Peter Kriegel, Paper ID 219.