ISSN XXXX XXXX © 2018 IJESC

Research Article Volume 8 Issue No.2

Web Crawler Design for Data Warehousing

Yugandhara .A. Thak[1], Gayatri .K. Machhirke[2], Priya .R. Meshram[3], Priyanka .S.Kale[4], Ravikant .M. Singh[5]
BE Student[1, 2, 3, 4, 5]
Department of Computer Technology
Priyadarshini College of Engineering, Nagpur, India

Abstract:
Usage of Internet is increasing day-by-day and with it size of data is also constantly increasing. The Internet is the largest collection of data today. Lot many web pages are available on the internet related to some keyword and to make specific search of any keyword we have made this crawler. A web crawler is also called as a Web Spider or a Web Robot. It is a program or automated script which browses the world wide web in a methodical, automated manner. This process is called web crawling or spidering. Many legitimate sites, in particular search engines use spidering as a means of providing up to date data. This will be making the search easy and fast. In the isolated area we can use this software as we can access the data in the offline mode. We can use this application in the rural area where we do not get the internet connection but there's one condition that we should have searched the data earlier.

I. INTRODUCTION

In our application we use the web crawler for data- warehousing which is generally used by the search engine. If we want to search anything from data-warehouse then by this application we can do it. A web crawler is program that visits websites and read their pages and other information in order to create entire search for a search engine index. The major search engines on the web all have such a program which is also known as a"spider" or a "bot". Crawlers are typically programmed to visit sites that have been submitted by their owners as a new or updated. Entire sites or specific pages can be selectively visited and indexed .Crawlers apparently gained the name because they crawl through a site a page at a time, following the links to other pages on the site until all pages have been read. In this application, user needs to tell target first and provide a domain with it specify which parts of the page one needs to crawl. Or either provides a list of direct URLs, and the application will extract data for the user. Now it's turn of machine learning to take over the hard work. Using advance scraping techniques, we crawl websites and extract raw data. All necessary metadata is saved for future references. If raw data is not enough, we can remove all unnecessary components and structures it to fit your needs. Once your data is ready, you can download it through API.

II. WORKING

The basic procedure executed by any web crawling algorithm takes a list of seed URLs as its input and repeatedly executes the following steps

- Remove a URL from the URL list.
- Download the corresponding page.
- Check the relevancy of the page.
- Extract any links contained in it.
- Add these links back to the URL list.
- After all URLs are processed.
- Return the most relevant page.

III. LITERATURE REVIEW

The processing of the text of web pages in order to extract information can be expensive in terms of processor time. Consequently a distributed design is proposed in order to effectively use idle computing resources and to help information scientists avoid the need to employ dedicated equipment. A system developed using the model is examined and the advantages and limitations of the approach are discussed. This paper present different types of crawler design based on crawler4j in java. We use crawler4j because it's an open source web crawler for Java which provides a simple interface for crawling the Web. Using it, we can setup a multi-threaded web crawler in few minutes. The proposed system is not only design for data-warehouse but it can also be used by any SEO(Search Engine Optimization) user. 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 to provide fast searches. Also we are applying a methodology where we will be saving all searched data in the data-warehouse and will be accessing it when in use. It's in this case giving a big advantage of accessing the data in offline mode also we are interfacing computer with the smart phone so that we can make access in the smartphone too.

IV. PROPOSED SYSTEM

In this paper we are providing a 4 types crawler design which allows user to crawl on given website and gather the data found on the website pages analysis it and saved them in by sorting it by given keywords.

4.1 Basic Crawler

a. Number Of Crawlers
Shows the number of concurrent threads that should be initiated for crawling.

b.Storage path
Folder where intermediate crawl data is stored for further used in data warehouse. Folder where index data is stored. For each job

the crawler need to provide different folder. Same folder cannot be used until application is restarted because the files in the folder are locked by background threads which release only when the application is closed.

c. Politeness Delay

To Make sure that we don't send more than 1 request per second (1000 milliseconds between requests).

d. Max Depth of Crawling

We can set the maximum crawl depth here. The default value is -1 for unlimited depth

e. Max Page To Fetch

We can set the maximum number of pages to crawl. The default value is -1 for unlimited number of pages

f. Include Binary Content In Crawling

If we want crawler4j to crawl also binary data ?example: the contents of pdf, or the metadata of images etc

g. Resemble Crawling

This config parameter can be used to set the crawl to be resumable (meaning that you can resume the crawl from a previously interrupted/crashed crawl). Note: if you enable resuming feature and want to start a fresh crawl, you need to delete the contents of rootFolder manually.

h. Use proxy

We can also tell the crawler to use the proxy for targeted website

4 .2Image Crawler

A simple image crawler that downloads image content from the crawling domain and stores them in a folder. This crawler is useful if image data is required for data mining. In this crawler we have following configuration options.

a. Number of Crawlers

Shows the number of concurrent threads that should be initiated for crawling.

b.Storage path

Folder where intermediate crawl data is stored for further used in data warehouse. Folder where index data is stored. For each job the crawler need to provide different folder. Same folder cannot be used until application is restarted because the files in the folder are locked by background threads which release only when the application is closed.

c.Include Binary Content In Crawling

If we want crawler4j to crawl also binary data ?example: the contents of pdf, or the metadata of images etc. by default its true because we are dealing with images

d.Download Folder

Folder path where all the images are saved.

4.3Thread based Crawler:

The crawler will act as a controller to collect data/statistics from crawling threads. In this crawler design following are the configuration options are available for user

a.Number Of Crawlers

Shows the number of concurrent threads that should be initiated for crawling.

b.Storage path

Folder where intermediate crawl data is stored for further used in data warehouse. Folder where index data is stored. For each job the crawler need to provide different folder. Same folder cannot be used until application is restarted because the files in the folder are locked by background threads which release only when the application is closed.

c.Max Page To Fetch

We can set the maximum number of pages to crawl. The default value is -1 for unlimited number of pages

4.4Multiple Crawler:

This is crawler use full if you want to use two distinct crawlers to be run concurrently. For example, you might want to split your crawling into different domains and then take different crawling policies for each group. Each crawling controller can have its own configurations.

a)Storage Folder

Folder where intermediate crawl data is stored for further used in data warehouse. Folder where index data is stored. For each job the crawler need to provide different folder. Same folder cannot be used until application is restarted because the files in the folder are locked by background threads which release only when the application is closed.

b)Politeness Delay

To Make sure that we don't send more than 1 request per second (1000 milliseconds between requests).

c)Max Page To Fetch

Max page to fetch

d)Number of Crawler Controller

Number of Multiple crawler controller to run on multiple domains concurrently

IV.CONCLUSION:

Reducing the search time with relevant results is one of the major problems faced by search engines these days. There are a large number of algorithms used by Web Crawlers to increase their efficiency and the approaches discussed in this report are also a part of that group. Search will give more relevant results even though initially it may consume some amount of time. The algorithm can work efficiently in static as well as less dynamic environments where environment variables do notchange during the search. If the environment drastically changes during the search then the current process will not be able toproduce efficient result. This is one of the major weaknesses in the current algorithm that does not allow dynamic change in Heuristic approach based on drastic changes in the environment.

V. REFERENCES

[1]. Mini Singh Ahuja Dr Jatinder Singh BalVarnica,"Web Crawler: Extracting the Web Data", International Journal of Computer Trends and Technology (IJCTT) – volume 13 number 3 – July 2014

[2]. Pavalam, S. M., SV Kashmir Raja, Felix K. Akorli, and M. Jawahar, "A Survey of Web Crawler Algorithms," International Journal of Computer Science, vol. 8, iss. 6, no 1, Nov. 2011.

[3]. Component of web search system figure, Accessed July 25,2017. https://www. google.co.in /url?sa=i&rct=j &q=&esrc=s&source=images&cd=&ved=0ahUKEwionMns_aTVAhXLpo8KHaB0BPQQjRwIBw&url=https%3A%2F%2Fen.wikipedia.org%2Fwiki%2FWeb_crawler&psig=AFQjCNGV10hs3Dm9EBk_yJgFmXskFXXq6g&ust=1501090566478463

[4]. [What is crawler? – Definition from Whatls.com] http://searchmicroservices.techtarget.com/definition/crawler

[5]. ShaliniSharma, “Web Crawler”, International Journal of Advanced Research in Computer Science and Software Engineering- Volume 4, Issue 4, April 2014

[6]. Aviral Nigam, “Web Crawling Algorithms”, International Journal of Computer Science and Artificial Intelligence Sept. 2014, Vol. 4 Iss. 3, PP. 63-67

[7]. Sharma, Sandeep, and Ravinder Kumar, “Web-Crawling Approaches in Search Engines,” Technical Report, Thapar University, Patiala, 2008.