Design of Migrating Web Parallel Crawlers: A Novel Image Change Detection Approach

Deepak
Student of M. Tech, Department of Computer Science Engineering
Indus Institute of Engineering & Technology, Jind, Haryana, India

Deepika Mahal
Assistant Professor, Department of Computer Science Engineering
Indus Institute of Engineering & Technology, Jind, Haryana, India

Abstract- With the increase in size of web, search engine depends upon the Web Crawler to download and build index of million/billion of pages for efficient information retrieval when user interact through search interface. This paper will include the definition of Web Crawler, proposed solution, system architecture, crawler manager, indexer, web server, crawl frontier and database. Some advantages with the design of crawler with comparison of solutions provided by researchers.

Keywords – Crawler Manager, Crawl Frontier, Indexer, Web Server, Database.

I. Introduction

A Crawler is a program which is used for downloading web pages from World Wide Web (which is basically a collection of text documents, images, multimedia and other resources which are linked by HYPERLINKS and URLs) for web search engine (is a software system which is used to search for information on World Wide Web). Web crawler is also called as Spider, Web scatters, Automatic indexer, wanderers, Web robots, ants, bots. As the time is changing size of World Wide Web is also changing, in the recent years it has grown from thousand to billion. Due to this explosion in size, web search engine are becoming increasingly important as they are used for locating information. Basically web search engine depends upon web crawler to create and maintain web indices for web pages. A web crawler works by starting with set of URLs (which are also called seed URLS) which are stored in queue like data structure. It works by downloading pages associated with these URLs extract any HYPERLINKS present in page and iteratively downloading the web pages identified by these HYPERLINKS. Basically after extraction of URLs, it enqueue it. There are various strategies it can either use breadth first, depth first or various other strategies.

II. PROPOSED ALGORITHM

```
A NOVEL IMAGE CHANGE DETECTION()
1. Do Forever
2.  Begin
3.  URL selected by Crawler Manager.
4. Crawler Manager pass URL to one of Migrant.
6. Migrant Check whether the URL has already
   been Entered or not.
7. If URL already existing in Database
8. Then
9. Fetch images from URL and computes  SHA256.
```

```
10. If  current image SHA256 value matches
    any of stored SHA values for that Url
11. Then
12. No change in Database.
13. Else
14. Update the Database.
15. Else
16. Fetch images and compute SHA256 values
    and passed to Indexer.
17. Indexer perform Indexing on URL name and
    SHA values.
    End
```

III. PROPOSED WORKS

In this section method was proposed which derive a code for images to determine whether they have undergone a change or not. Ideally a change in a link to an image hyperlink will be reflected in the label of the hyperlink for that image and the same will be depicted by the formula proposed above. But in case the text does not change but the image is replaced, it will still be left undetected. We propose the following method for image change detection:

Image change detection is divided into three steps:

- First process is preprocessing of an inputted images to find SHA256 value.
- Second step includes, two inputted images are compared by SHA256 to determine the difference image.
- Third processing step which is to analysis of the difference image.

SYSTEM ARCHITECTURE

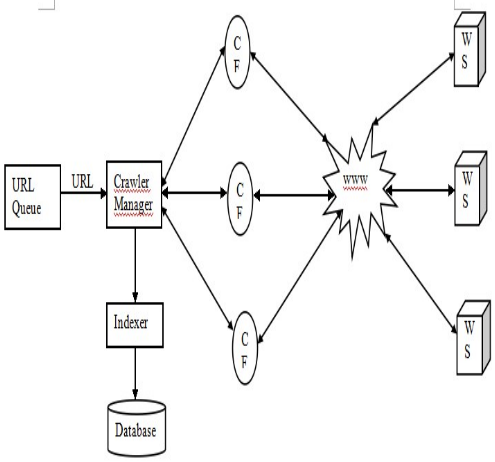

Figure1: Migrating Parallel Crawler With Web page Change Detection

- CRAWLER MANAGER: Central crawler is the central component of system and started first in system where all other components are started and register with central crawl manager to offer or request services. All URLs to be crawled are selected by central crawler. It is only component visible to other components while all crawl frontiers work in independent manner. It is responsibility of central crawler to query DNS resolver for IP addresses of host or servers, after getting the URL's of files. Also checks robots. text file in root directory of web servers. Crawl frontiers have to register themselves with central crawler and are logically migrated to different locations. Different URL's are assigned to all the crawl frontiers and then after receiving URL they started crawling. In this way the compressed downloaded content from crawl frontiers are passed to central crawler. It also passed the data to indexer for managing the database.

- CRAWL FRONTIERS: All the processing is done by the migrants. Migrants check the URL whether it is first time or it has been entered earlier. If the URL is entered first time then migrant fetch the images from that URL and compute there SHA2556 values and passed to crawler manager. But if URL is not entered first time then migrant fetch the images and compute there SHA256 values and compare with SHA values stored for that URL. If all the values are same then there is not change otherwise migrant pass the changed images to crawler manager. Migrants passed the data to crawler manager in compressed form and save the network bandwidth.
- WEB SERVER: A web server is a program that serves the files form web pages to web users. It uses World Wide Web's hypertext transfer protocol (HTTP) and client/server model.
- INDEXER: Crawler Manager passed the data to Indexer for indexing. Indexing is performed on URL name and SHA256 values.
- DATABASE: Shared database is used i.e access is given to crawler manager and all the migrants. Database maintain URL names, stored images names and there SHA256 values and also the time of visit to that URL.

FLOWCHART

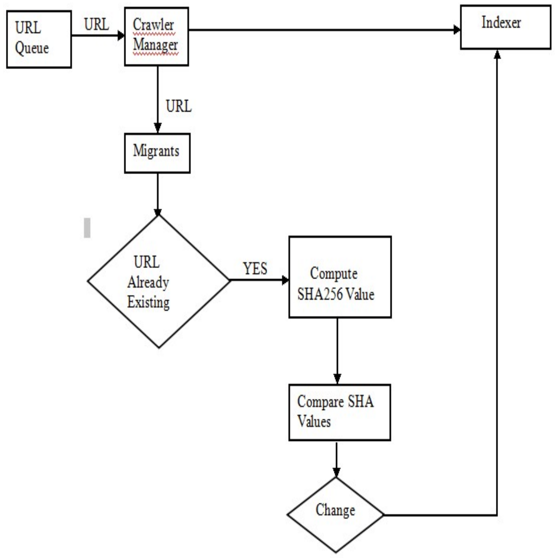

Figure 2: Flowchart for Image Change Detection

- Fig depicts process flow for proposed system to change detection.
- First URL is selected by Crawler Manager and passed to any of migrant.
- Migrant then check the URL whether it is first time or it has been entered earlier. If the URL is entered first time then migrant fetch the images from that URL and compute there SHA2556 values and passed to crawler manager.
- But if URL is not entered first time then migrant fetch the images and compute there SHA256 values and compare with SHA values stored for that URL. If all the values are same then there is no change otherwise migrant pass the changed images to crawler manager.

- Then Indexer start indexing on the basis of URL name and SHA256 value and maintain the updated database.

IV. ADVANTAGES

- New Product Uploaded : This will help the users to came to know about if new product is uploaded on the website and will show the image of that product. So that user can get clear information about the product.

- Design change: this is also useful to tell if there is any change in the design of the website. For example consider the case that the logo of a website have been changed then this can easily detect that change.

- Frequency Set: In this frequency of revisiting at a website can be set. It can be set at periodic hours.
- Maintain History: It stores the last updated time of a URL in the database. According to these update time different URL's can be categorized i.e which URL need to be revisited very frequently and which URL can be revisited after a long time.

- Less Collision: in this we are using SHA256 hash value for images. Chances of collision are very few. So it increases the efficiency.

- Very Fast: This method of detecting the changes is very fast because it saves lots of network bandwidth and hence increases network efficiency.

- Banner Change: This feature is more useful in the case of online shopping websites because users can quickly came to about new banner uploaded.

- Image seen earlier or not : it can also detect whether a particular image has been uploaded for the very first time or it has been uploaded earlier.

V. CONCLUSION

In this Paper, we have described the architecture and implementation details of our crawling system, and presented some preliminary experiments. There are obviously many improvements to the system that can be made. A major open issue for future work is a detailed study of the scalability of the system and the behaviour of its components. This could probably be best done by setting up a simulation testbed, consisting of several workstations, that simulates the web using either artificially generated pages or a stored partial snapshot of the web. We are currently considering this, and are also looking at testbeds for other high-performance networked systems. This future work will deal with the problem of quick searching and downloading the data. The data will be collected and analyzed with the help of tables and graphs.

REFERENCES

[1] Navita, Mahesh “A Novel Architecture for Topic specific web crawler”.

[2] Junghoo Cho ,Hector Garcia-Molina, “Parallel Crawlers”,In proceedings of WWW2002, Honolulu, hawaii, USA, May 7-11, 2002. ACM 1-58113-449-5/02/005.

[3] Douglas E. Comer, “The Internet Book”, Prentice Hall of India, New Delhi, 2001. [2] Monica Peshave, “How Search Engines Work And A Web Crawler Application”

[4] S. S. Vishwakarma, A. Jain and A. K. Sachan, “A Novel Web Crawler Algorithm on Query based Approach with Increases Efficiency”, published in International Journal of Computer Applications (0975 – 8887), vol. 46, no. 1, (2012) May.

[5] Ashutosh Dixit,U.C. Pant, “A Novel and Efficient Mechanism for Securing Migrating Crawler Data”.

[6] A.K. Sharma, Ashutosh Dixit, Deepika, and Niraj Singhal, “Security issues in Distributed Crawling based on Mobile Agents”, 5th International IT Conference on Downtrend Challenges in IT, DCIT09, Ludhiana, Punjab, May 2009.

[7] A.K.Sharma, J.P.Gupta, , D.P.Agarwal, “ An architecture of parallel crawler based on Augmented Hypertext documents “

[8] Dr Rajender Nath, Khyati Chopra “Web Crawlers: Taxonomy, issues and challenges”.

[9] Md. Faizan Farooqui1, Dr. Md. Rizwan Beg, Dr. Md. Qasim Rafiq “An Extended model for Effective migrating parallel Web crawling with domain specific and Incremental crawling“