ISSN:0975-9646

Investigations on the Effect of Various Parameters on the Offline Browsing Efficiency of a Web Archiving System Using Breadth First Search Algorithm

B.Vijaya Babu[1], P.Sunitha[2],
[1]Department of Computer Science and Engineering, K L University
GUNTUR , Andhra Pradesh, INDIA

[2]Department of Computer Science and Engineering,
Dhanekula Institute of Engineering and Technology,
Ganguru, VIJAYAWADA , Andhra Pradesh ,INDIA

Abstract— **In this paper, the effect of various parameters on the offline browsing efficiency of a web archiving system using breadth first search Algorithm is investigated. A web crawler based multithreaded web archiving system is designed using breadth first search algorithm and the offline browsing efficiency of the web archiving system is estimated in the presence of the parameters like the searching algorithm, browsing tool , speed of the Internet connectivity and the processing system configuration.**

Keywords— **Digital preservation, Offline browsing efficiency, breadth first search algorithm, multithreaded web archiving system.**

I. Introduction to web archiving

Web Archiving (or mirroring) is a technique aimed at downloading the web pages of a website successfully on to a user specified location and these downloaded pages can be browsed at a later time during the offline conditions. This process re-creates the entire web site as the mirror of the original web site at user specified location. The process of downloading and saving the data present on a particular website on to the local storage media facilitates the digital preservation of the web data .Digital preservation of the websites can be done by saving each and every page and the related links of the site individually. Since the website consists of thousands of web pages and other related links, the digital preservation needs a lot of time and effort. Availability of proper Internet connectivity also plays an important role for the digital preservation. But, due to various technical reasons, availability of proper Internet connectivity always may not be expected. In these circumstances there should be a mechanism that not only preserves the digital data, but also browsing this data when the Internet connectivity is not present (i.e. during offline conditions).[6][7][8]

II. Literature survey

R. Baeza-Yates et al. [14] presented the trade-offs in designing efficient caching systems for web search engines. They have explored the impact of different approaches, such as static vs. dynamic caching, and caching query results vs. caching posting lists. They have proposed a new algorithm for static caching of posting lists, which outperforms previous methods. Masanes.J[15][16][17] presented various crawling algorithms and approaches undertaken today by different institutions; it will discuss their focuses, strengths, and limits, as well as a model for appraisal and identifying potential complementary aspects amongst them. Rui Cai et al. [15] discussed about an intelligent crawler with the main idea of learning about the site map of a forum site with a few pre-sampled pages, and then decide how to select an optimal traversal path to avoid duplicates and invalids. M. Angeles Serrano et al. [17] reported a detailed statistical analysis of the topological properties of four different WWW graphs obtained with different crawlers. They have studied the statistical measures beyond the degree distribution, such as degree-degree correlation functions or the statistics of reciprocal connections B. J. Jansen et al. [18] provided a classification for information agent using stages of information gathering, gathering approaches, and agent architecture**.** Junghoo Cho et al. [16] studied the order in which a crawler should visit the URLs it has seen, in order to obtain more important pages first.

III. Design methodology

The basic structure of the World Wide Web (WWW) can be viewed as a directed Graph. The Pages and hyperlinks of the websites may be viewed as nodes and edges in a directed graph. This Web as a graph has billions of nodes and appears to grow exponentially with time. All most all of the search engines and archiving applications and related tools use web crawlers as the key component for the downloading and archiving process. Due to the commercial as well as business reasons and related competitive issues, they keep the internal design as trade secrets and used as copy righted materials. There fore they won’t reveal the design and architectural details to the public in general. More over, they keep the algorithms used as confidential and change them more frequently as part of their research & development activity, in order to prevent others so that they can’t change their data bases. The scalability issues of such applications will also be changed from company to company who develops these applications. The web archiving system is designed for offline browsing using **Breadth First** graph searching algorithms. It is also designed with multiple robots with preemptive multithreading as it allows the operating system to determine when a context switch should occur. More over, using multiple robots with proper synchronization, an increase in the rate of download of URLs/pages is expected. A provision of customized input active threads and option of selecting the graph searching algorithm are included in the system.

IIIA. Algorithm for Best First Search in Web Archiving System

Step 1

Start

Step 2

Go to the package workBench and implement the Method crawlerEditor ()

Create a Reference variable searchOrderChoice to the Choice interface and call addItem () methods

Step 3

Implement the addItem () with BFS as a Argument.

```
vertexList[0].wasVisited = true; // mark it
  displayVertex(0);              // display it
  theQueue.insert(0);            // insert at tail
  int v2;
  while( !theQueue.isEmpty() )     // until queue empty,
    {
    int v1 = theQueue.remove();   // remove vertex at head
    // until it has no unvisited neighbors
    while( (v2=getAdjUnvisitedVertex(v1)) != -1 )
      {                         // get one,
      vertexList[v2].wasVisited = true;  // mark it
      displayVertex(v2);              // display it
      theQueue.insert(v2);            // insert it
      }   // end while
    }  // end while(queue not empty)
  // queue is empty, so we're done
  for(int j=0; j<nVerts; j++)           // reset flags
    vertexList[j].wasVisited = false;
  }  // end bfs()
// -------------------------------------------------------------
  // returns an unvisited vertex adj to v
  public int getAdjUnvisitedVertex(int v)
    {
  for(int j=0; j<nVerts; j++)
    if(adjMat[v][j]==1 && vertexList[j].wasVisited==false)
      return j;
  return -1;
    }
  }
```

Step 4

Select the url fields, depth fields and searchOrderChoice using handleEvent () method.

Step 5

Select the hyper links ,images and all links using setCrawler() method

Step 6

Get the domains, links and other urls using getCrawler() method.

Step 7

Repeat Step 3 to 6 until selectIndex==0.

Step 8

Stop.

IV. SIMULATION RESULTS

To implement the web crawler based archiving system, the Workbench [19][20] tool is used. Workbench is a general-purpose crawling tool that runs as a Java applet inside a Web browser. The Workbench provides the user with an opportunity to make the archiving application more adaptable and tailored, and also several graphical visualizations to gauge the effectiveness of the application and improve it iteratively. Running the archiving Workbench inside a browser provides a fair level of integration between browsing and archiving, so that archiving tasks that arise during browsing can be solved immediately without changing context.

The archiving system is implemented for the web site www.jnettechnologies.com. The archiving system is run on the system whose hardware configuration with Intel Core I3 Processor having processor speed of 2.27GHz, 4GB RAM, 32 bit Operating System. The Windows 7 is used as the operating system. The Internet Explorer 8.0 Version Browser is used for browsing the Web. The browsing speed of the Internet connection (Wi-fi) is 10MBPS and BSNL as the service provider.

The website www.jnettechonologies.com has 122 correct internal URLs as per the MIME standards. This was verified by Xenu Link Sleuth 1.3.8 version tool that gives the information about all types of links and URLs present in a particular website.

V. DISCUSSION OF THE RESULTS

A. The Offline Browsing Efficiency

The offline browsing efficiency is defined the percentage of the ratio of number of URLs or pages retrieved by the archiving system to the total number of URLs or pages that the website originally possesses.

$$\text{Offline Browsing Efficiency} = \frac{\text{Number of URLs or pages retrieved by the system X 100}}{\text{Total number of URLs (or pages), the website originally possesses.}}$$

The off line browsing efficiency may mainly depend on

i) Searching algorithm used in the archiving system

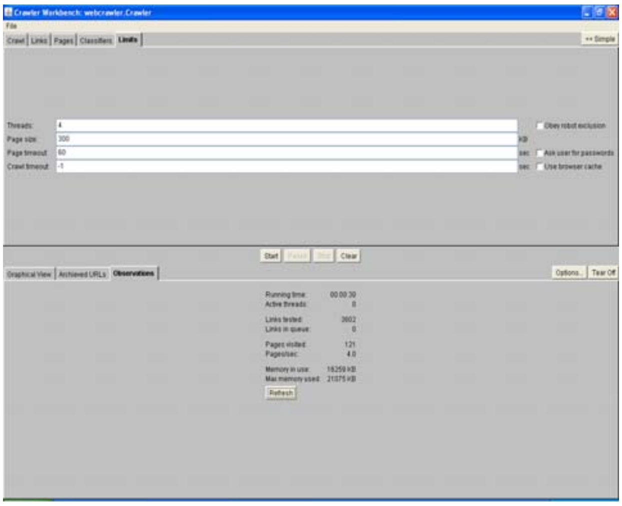
ii) Browsing tool used to browse the website

iii) Design technology of the websites that follows the standards of naming conventions and other file name extensions

iv) Estimation of Offline Browsing Efficiency Using Breadth First Search Algorithm.

The system is implemented for the website www.jnettechnologies.com using Breadth First Search Algorithm for different active thread connections starting from 0 to 12.

A Screenshot of the Breadth First Search Algorithm for the website www.jnettechnologies.com is shown in figure 1.



Threads: 4 □ Obey robot exclusion
Page size: 300 KB
Page timeout: 60 sec □ Ask user for passwords
Crawl timeout: -1 sec □ Use browser cache

Running time: 00:00:30
Active threads: 0

Links tested: 3602
Links in queue: 0

Pages visited: 121
Pages/sec: 4.0

Memory in use: 16259 KB
Max memory used: 21075 KB

Fig. 1 Screen Shot of Web Archiving System for Active threads=4 using Breadth First Search Algorithm

The observations of the variations of pages/URLs visited for various active threads in Breadth First Search algorithm are tabulated as shown in table 1.

TABLE 1
Active Threads Vs Pages/URLs Visited

Breadth First Search (BFS) Algorithm	
Active Threads	**Pages/URLs Visited**
0	**0**
1	**121**
2	**121**
4	**121**
6	**121**
8	**121**
10	**121**
12	**121**

The variation of URLs/Pages visited for various active threads is shown in the figure2 below.

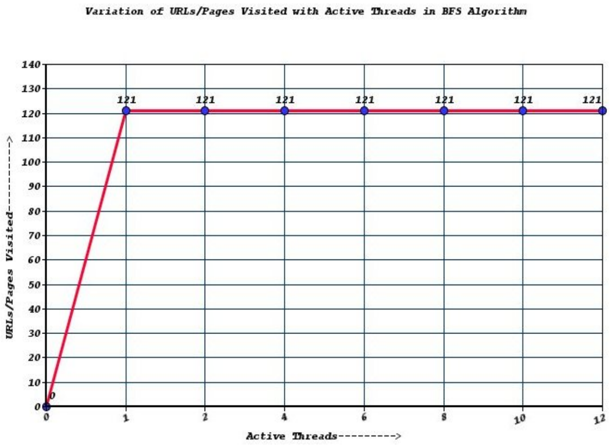


Figure 2. Graph showing the variation of URLs/Pages Visited with Active Threads in Breadth First Search Algorithm

On an Average the archiving system retrieves 105.875 URLs/Pages.

Now, the Percentage of offline browsing efficiency in this case is given by 105.875*100/122=86.782 %

CONCLUSIONS

The off line browsing efficiency mainly depends on Searching algorithm used in the archiving system, browsing tool used to browse the website, design technology of the websites that follows the standards of naming conventions and other file name extensions. Based on these factors, different case studies were conducted on various Internet connectivity environments with varying speeds and bandwidths for different active threads. Even though the running time for the entire website to be archived mainly depends on the speed and the bandwidth of the network connectivity it was observed that as the number of active threads increases, the running time decreases and vice versa. But it was also observed that using more robots or multiple active thread connections increases the rate of downloading of pages up to a certain point, and once bandwidth saturates the system is not considering the CPU processing time and adding more threads increases the performance monotonically.

REFERENCES

[1] Masanès, J: Web Archiving, *Berlin: Springer-Verlag. ISBN 3-540-23338-5, 2006.*

[2]. Filip Boudrezand Sofie Van den Eynde: Archiving Website , *DAVID ,2002*

[3] Brügger, N :Archiving Websites. General Considerations and Strategies, *The Centre for Internet Research. ISBN 87-990507-0-6, 2005.*

[4] Nina Tahmasebi, Sukriti Ramesh and Thomas Risse : First Results on Detecting Term Evolutions **,** *In the Proceedings of the 9th International Web Archiving Workshop (IWAW 2009) Corfu, Greece, September/October, 2009*

[5] Day, M. Preserving the Fabric of Our Lives: A Survey of Web Preservation Initiatives, *Research and Advanced Technology for Digital Libraries: Proceedings of the 7th European Conference (ECDL): 461–472, 2003.*

[6] Eysenbach, G. and Trudel, M.: Going, going, still there: using the WebCite service to permanently archive cited web pages, Journal *of Medical Internet Research, vol* 7, 2005

[7] James Lee: Database Archiving: A Critical Component of Information Lifecycle Management, *Data Base Journal, April 21, 2004.*

[8] Lyman, P : Archiving the World Wide Web: *Building a National Strategy for Preservation: Issues in Digital Media Archiving*, 2002

[9] Andrei Broder, Ravi Kumar, Farzin Maghoul, Prabhakar Raghavan, Sridhar Rajagopalan, Raymie Stata, Andrew Tomkins, and Janet Wiener : Graph structure in the web: experiments and models, *In the Proceedings of the Ninth International World-Wide Web Conference, Amsterdam, Netherlands, May 2000.*

[10] Sergey Brin and Lawrence Page: The anatomy of a large-scale hyper textual web search engine, *In the Proceedings of the Seventh International World-Wide Web Conference, Brisbane, Australia, April 1998.*

[11] Junghoo Cho and Hector Garcia-Molina. The evolution of the web and implications for an incremental crawler. *In the Proceedings of the Twenty-sixth International Conference on Very Large Databases*, Cairo, Egypt, September 2000.

[12] Filippo Menczer, Gautam Pant, and Miguel E. Ruiz: Evaluating topic-driven web crawlers. *In the Proceedings of the Twenty-Fourth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New Orleans, LA, September 2001.*

[13] Mircea-Dan Antonescu, Mark Guttenbrunner and Andreas Rauber : Documenting a Virtual World - A Case Study in Preserving Scenes from Second Life, *In the Proceedings of the 9th International Web Archiving Workshop (IWAW 2009) Corfu, Greece, September/October, 2009.*

[14] Myriam Ben Saad, Stéphane Gançarski and Zeynep Pehlivan :A Novel Web Archiving Approach based on Visual Pages Analysis , *In the Proceedings of the 9th International Web ArchivingWorkshop (IWAW 2009) Corfu,Greece,September/October,* 2009

[15] Rui Cai, Jiang-Ming Yang, Wei Lai, Yida Wang, and Lei Zhang : iRobot: An Intelligent Crawler for Web Forums , *In the proceedings of WWW, Beijing, China 2008 .*

.[16] Junghoo Cho, Hector Garcia-Molina, and Lawrence Page: Efficient crawling through URL ordering. *In the Proceedings of the Seventh International World- Wide Web Conference, Brisbane, Australia, April 1998.*

[17] M.A'ngeles Serrano, Ana Maguitman, Mari ´ An Bogu˜n, Santo Fortunato and Alessandro Vespignani : Decoding the Structure of the WWW: A Comparative Analysis of Web Crawls, *In the ACM Transactions on the Web, Vol. 1, No. 2, Article 10, August 2007.*

[18] Bernard J. Jansen, Tracy Mullen, Amanda Spink, and Jan Pedersen : Automated Gathering of Web Information: An In-Depth Examination of Agents Interacting with Search Engines , *In the ACM Transactions on Internet Technology, Vol. 6, No. 4, Pages 442–464, November 2006* .

[19] www.cs.cmu.edu/~rcm/websphinx

[20] Robert C. Miller and Krishna Bharat : SPHINX: A Framework for Creating Personal, Site-Specific Web Crawlers, *In Proceedings of WWW7, Brisbane Australia, April 1998.*

Dr B.Vijaya Babu is presently working as Professor in CSE department of K L University, Vaddeswaram, Guntur(D.t) Andhra Pradesh,INDIA. He has obtained B.Tech.,(ECE) degree from JNTU College of Engineering,KAKINADA, M.Tech.,(CSE) degree from JNTU College of Engineering,ANANTAPUR and PhD degree from Andhra University,VISAKHAPATNAM .He has published several research papers in various International journals and attended International Conferences conducted in India.

Ms P.Sunitha is presently working as Assistant Professor in CSE department, Dhanekula Institute of Engineering & Technology, Ganuru,VIJAYAWAD, INDIA. She obtained her B.Tech.,(ECE) degree from JNTU, HYDERABAD and M.Tech.,(CSE) degree from JNTU,KAKINADA.She has published several research papers in various International journals and attended International Conferences in India.