Advances in Computer Science and Information Technology (ACSIT)
Print ISSN: 2393-9907; Online ISSN: 2393-9915; Volume 2, Number 3; January-March, 2015 pp. 200-203

http://www.krishisanskriti.org/acsit.html

Reduction of Load on Network using Compression Technique with Web Crawler

Sneha Tuteja[1], Naresh Kumar[2] and Rajender Nath[3]

[1]B. Tech. Student MSIT, Janakpuri, New Delhi
[2]CSE Dept., MSIT Janakpuri, New Delhi
[3]DCSA Kurukshetra University Kurukshetra

Abstract—*The following research paper aims to present an algorithm that would reduce the ever increasing load on the network. This research paper tries to incorporate an efficient compression technique along with a Web crawler that would compress the downloaded Web page at remote side. This compressed Web page would then be transferred to search engine's end for storage and indexing. The paper has made use of the already proposed algorithms by certain authors and has modified them in order to provide a better and an efficient algorithm.*

1. INTRODUCTION

Thousands of people use Internet on a day to day basis [1]. Internet has become a essential part of many people's life. Moreover, its popularity and efficiency is increasing day by day [2]. But at the same time, various network related issues come forward which not only degrades the effectiveness of the search engine's result but also is a reason of many problems that are faced by users. Various network related issues are listed below:-

a) Due to the dynamic nature of the Web, a Web page changes even before the previous version could be downloaded. Thus, every time a Web crawler downloads the Web page and transfers it to the search engine's database, the novelty of the Web page diminishes.
b) Since there is not only a single search engine that is in use and every search employs several Web crawlers [3]. According to an estimate, these hundreds of Web crawlers are responsible for about 60% of the total network's traffic.
c) Due to the large amount of network load, user faces end to end transmission delay. Moreover network traffic is responsible for speed reduction while transmitting content from one site to another [4].

This paper modifies an already existing algorithm in order to reduce the bandwidth consumption and the network load by incorporating compression technique that gives compression percentage up to 35-40% with the Web crawlers [5]. This would reduce the amount of data that has to be transferred via a network to the search engine's end and would finally reduce the usage of resources.

The remaining research paper has been composed as follows: - Section 2 gives the related work which is followed by the problem formulation in section 3. Section 4 discusses the proposed work for reduction of load. Section 5 provides the various analysis of the proposed algorithm. And finally section 6 concludes the paper.

2. RELATED WORK

A lot of research has been done to reduce the amount of load present on the network. In [6], Joachim Hammer and Jan Fiedler proposed an approach in which they allow the Web crawlers to control the granularity of the Web documents. They let the Web crawler to choose the relevancy of the document as well as the information that is presented in it. Search engine transfers the crawler to the remote site where it lowers the amount of information that it feels is not relevant to the desired query. Finally, it transfers only that information which it considers important for answering the user's query. The drawback of such an approach is that the Web crawler is actually ruining the information that has been presented by the Web document. Moreover, it is not possible for the Web crawlers to decide whether the information presented in Web page is important for the users or not.

Rajender Nath and Satinder Bal in [7], presented a way which compared the original pages with the changed pages on the basis of the keywords and the URLs. The original page has already been downloaded and stored in the old database of the search engine. This comparison tells whether the Web page has been revised or not. If the Web page has not been modified since the last crawl, the Web crawler would not use the bandwidth and various other already limited resources to download and transfer the same page via a network. This results in the better utilization of available web resources. The authors selected 100 Web pages for experiment. By the use of proposed approach they found that only 63% of them have

been modified. Based on this refreshment of pages and weight of the additional component they calculated the load on the network.

The approach is further modified by Naresh Kumar and Rajender Nath in [8]. They proposed a way to calculate the change frequency of a Web page. They used a frequency change estimator to verify whether a page has been changed or not.

Further, there is comparator module that compares the old pages with the new ones on the basis of ASCII count. If the ASCII counts of the old Web page matches with the ASCII count of the new page, the Web page has not been modified and is left by the crawler. Otherwise, the Web crawler downloads the new page replacing the older one in the search engine's database. In this only the ASCII value has been taken by the Web crawlers to the remote side and not the complete Web page for comparison. This would reduce the weight of the additional component by a large amount. The authors had found total 60 pages out of 100 pages that have been modified since the last crawl. Moreover, the weight of the additional component is also reduced as only the ASCII value has to be sent along with the crawlers to the remote side. If this weight has been assumed to be 20Kb then the total load on the network has been calculated as 500kb.

3. . PROBLEM FORMULATIONS

There are certain problems that have been found in various techniques that were proposed by different authors. Some of these problems have been expressed below:-

a) Certain papers [7][8] that uses a page change calculating technique has to employee an additional component that has to be transferred with the Web crawlers to the remote side. Sending a complete Web page for comparison would increase the inefficiency of the proposed algorithm as a lot of bandwidth would be consumed by the Web crawlers and the remote side may not allow crawlers to utilize much of its memory.
b) Even after page change calculating technique is used, the reduction of load is not substantial. Therefore a significant compression technique can result in a significant reduction of load on the network.
c) Moreover, some of the compression techniques [9] are lossy and would change the information or the data that is meant to be provided by the Web page.
d) Another problem that can be mentioned here is that the proposed technique in [7][8] and many other does not provide the adequate information about how many times the respective Web page has been changed.

Main concern of the following research paper is to rectify first three problems that are stated above.

4. PROPOSED WORK

The authors had tried to find the compression technique that certainly gives a higher compression ratio so that by incorporating it with the Web crawlers maximum compression ration ratio is achieved which would reduce the amount of data that has to be transferred to the search engine. Compression ratio is defined as the ratio of uncompressed size to the compressed size of the Web page more the compression ratio, better it would be. For this authors have surveyed various compression techniques that are proved to be highly efficient such as WinZip, WinRar, GZip, 7zip etc. The compression ratio for the same is shown in Fig. 1.

Following are the results of the comparison that has been done by surveying various sites [10]:-

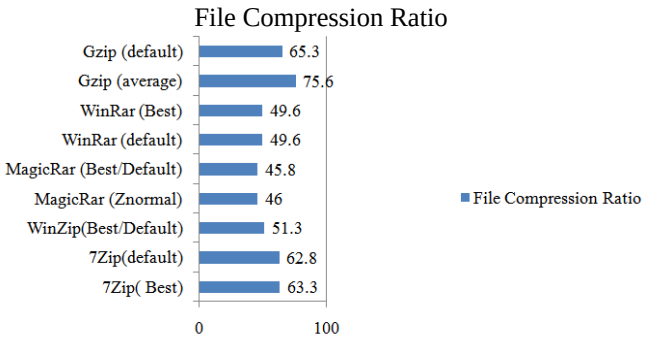

Fig. 1: compression ratios of various compression techniques.

By the above graph, it can be concluded that GZip gives the best possible compression ratio. Moreover, authors have practically tested the GZip compression technique as shown in Fig. 2 and the results are quite similar to the compression ratio mentioned above.

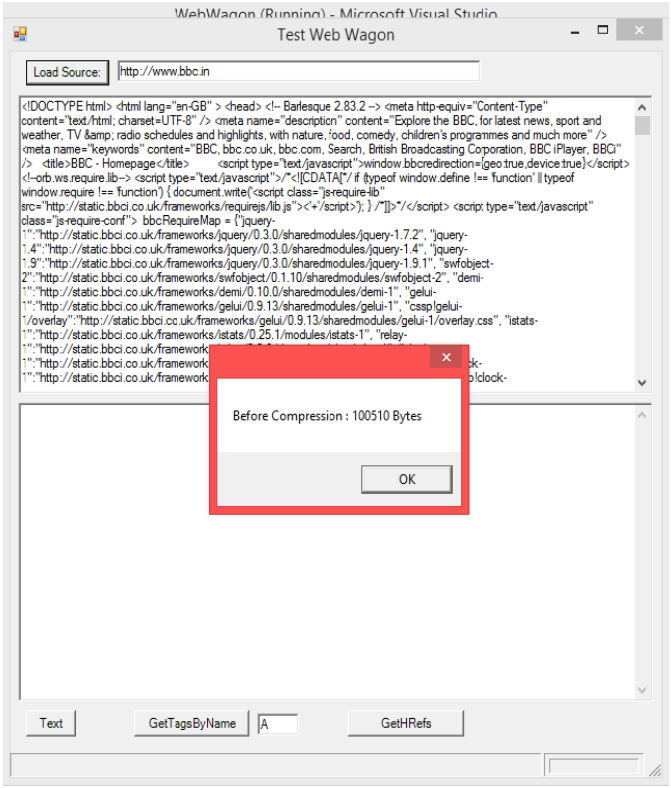

Fig. 2: Size of the Web page ‘ http://bbc.in’ in bytes

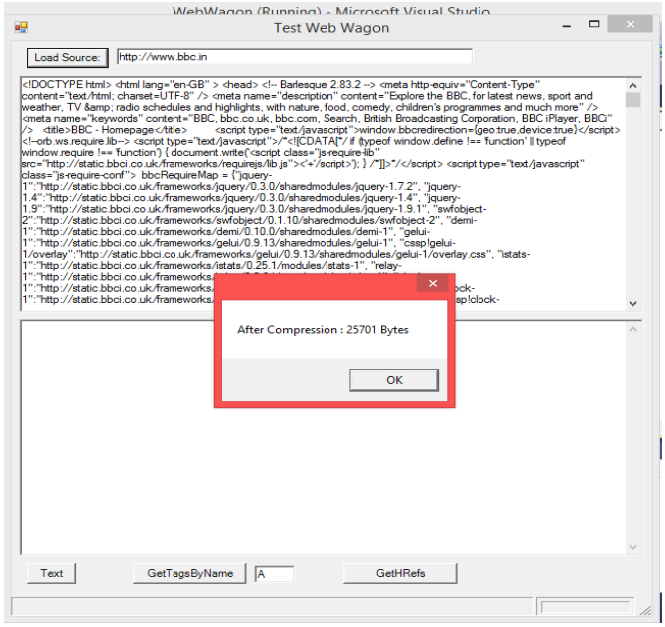

Fig. 3: Compressed size of the webpage ‘http:// www.bbc.in’ in bytes.

The size of the webpage ‘http:// www.bbc.in’ is 100510 in bytes. But when the same Web page is compressed by using the GZip then the result is shown in Fig. 3.

Compressed size of the same page is 25701 bytes.

After calculating, the compression percentage by the formula

Compression percentage= (uncompressed size - compressed size)/ uncompressed size

The percentage turns out to be 74.5.

Altering the architecture that is proposed by Satinder Pal and Rajender Nath [7] and further modified by Naresh Kumar and Rajender Nath, authors have replaced the WinZip technique that has been used by the authors with the GZip compression technique which is evidently more efficient in terms of compression ratio and time of usage.

5. ANALYSIS OF PROPOSED ALGORITHM

The given algorithm has the following terminology:-

Pt: Number of pages that are changed since the last crawl

Av: Average size of a page

Cm: Compression percentage

Wc: Weight of the extra component

Now, for calculating the load on the network, the following formula has been proposed and is shown in (1).

$$\text{Load} = (Pt * Av)\,(1 - Cm/100) + Wc \qquad (1)$$

The above formula gives the load on the network that has been put due to the Web crawlers. Wc represents the weight of the extra component that is taken by the Web crawlers on the remote side in order to check if the page has been modified or not. And Cm is the compression percentage of the technique that is being used by the search engine’s end and has been incorporated with the Web crawlers.

For verifying, the reduced load using WinZip and GZip techniques individually; and a traditional crawler that doesn’t utilizes any page change technique or any compression technique has been calculated.

Average size of a Web page has been assumed as 8 kb whereas the size of the extra component that is taken along with the Web crawlers to the remote side is assumed to be as 20 kb.

For traditional crawlers, it is assumed that there are a total of 100 pages. Since traditional crawlers do not use any page change calculating technique or compression technique, the load on the network as calculated by equation 1 represented above is as follows:-

100*8 = 800 Kb

For a Web crawler that uses a page change calculating technique [8], found that only 60% of the total pages have been modified and are needed to be downloaded.

Therefore Pt would be 60 for the above scenario and the load without compression by using the same equation would be calculated as follows:-

(60*8) +20 =500 Kb

By using WinZip compression that could compress the page up to 60 %, load will be:-

(60*8) (1- 60/100) +20 = 212Kb

And using GZip the load can be further reduced to slightly higher compression percentage [11]. The load on the network would be:-

(60*8) (1-70/100) +20 =164kb

It has been demonstrated that by using GZip the load on the network is less as compared to the scenario when WinZip compression technique is used.

6. CONCLUSION

This paper aims at developing a better approach in order to reduce the load on the network. The authors analyze various techniques for network load reduction by using available online services. Then they found GZip as a superior technique. To cross verify the results provided by online resources, the authors tested the GZip practically. Then the achieved results showed that the proposed technique reduced the network load from 500Kb to 164Kb which is much better than the network load mention in the available literatures. The most important task of this work is to implement the proposed approach in combination with mobile agents for network load reduction up to a significant amount. Several interesting issues may be identified which need to be addressed by future research before the implementation of proposed approach.

REFERENCES

[1] URL: http://www.Internetworldstats.com/stats.htm.

[2] URL: http://worldwideWebsize.com.

[3] S. Lawrence and C. Lee Giles, “Accessibility of information on the Web”, in Nature, 400(6740):107–109, July 1999.

[4] Google Inc. Google, September 2003. URL: http://www.google.com.

[5] URL: https://developers.google.com/Web/fundamentals/performance/optimizing-content-efficiency/optimize-encoding-and-transfer.

[6] Jan Fiedler and Joachim Hammer, “Using the Web Efficiently: Mobile Crawlers”, In Proceedings of the Seventeenth AoM/IAoM International Conference on Computer Science, San Diego, CA, 1999.

[7] Dr. Rajender Nath and Satinder Bal, “A Novel Mobile Crawler System Based on Filtering off Non-Modified Pages for Reducing Load on the Network”, in The International Arab Journal of Information Technology, Volume 8, No. 3, July 2011.

[8] Dr. Rajender Nath and Naresh Kumar, “A Novel Parallel Domain Focused Crawler for Reduction in Load on the Network”, in International Journal of Computational Engineering Research, Volume 2, Issue. 7, ISSN 2250-3005(online), pp.77-84, November 2012.

[9] URL: http://en.wikipedia.org/wiki/Lossy_compression.

[10] URL: http://www.tomshardware.com/reviews/winrar-winzip-7-zip-magicrar,3436-7.html.

[11] URL: http://www.differencebetween.net/technology/difference-between-zip-and-gzip.