

An Efficient Algorithm for Webpage Change Detection For Network Load Reduction

Ms. Kompal

Assistant Professor

Abstract: As the size of the web continues to grow, searching it for useful information has become increasingly difficult. Sufficient of current Internet traffic and bandwidth consumption are due to the web crawlers that retrieve pages for indexing by the different search engines. These crawlers also cause load on the remote server by using its CPU cycles and memory.

Moreover, due to the dynamic nature of the web, it becomes very difficult for a search engine to provide fresh information to the user. The web pages are uploaded and updated very frequently. There is a frequent change in the content of the web page hence it become necessary to develop an efficient system which could detect these changes efficiently and in the minimum browsing time. The mobile crawlers filter out pages that are not modified since the last crawl before sending them to the search engine for indexing purpose and to achieve this the old web page is compared with the new web page. The web page change detection system can be implemented by using various Tools or Algorithms. In this paper a new algorithm is presented to detect content level changes. It uses the text Code to detect content level changes by using the character digit position in the word.

Keywords: Mobile crawler, Network load, Pagechange, Search engine.

1. INTRODUCTION

The Search Engine creates and maintains data base for indexing purpose to process search related queries. The Search Engines internally use web crawlers to keep the fresh copies of data . A Web Crawler searches through all the Web Servers to find information about a particular topic. The crawler can only download a fraction of the Web pages within a given time. A user may visit certain Web pages repeatedly and is interested to know how each document has changed since the last visit. Due to rapid changes in the content of the web pages it has become very necessary to develop a Web Page Detection System which can detect these changes in minimum browsing time. It allows the user to find the items in web page which change frequently.

The crawling strategy must be able to decide about page change [1] without completely analyzing the page. This paper proposes an efficient method which intelligently decide about changes in a page using various parameters and skip unchanged pages.

Web page changes can be classified as follows:

- Structural Changes
- Content Or Semantic Changes
- Presentation Or Cosmetic changes
- Behavioural Changes

Structural Changes occur whenever a tag or link is added or deleted in a web page i.e. addition or deletion of a tag causes structural change [3] in a web page. Content or Semantic Changes occur whenever the content of a web page changes according to the reader point of view [3].

Presentation or Cosmetic Change Occur whenever the appearance of a web page is modified but the content of a web page remains the same [3]. Behavioural changes refer to changes in the active components which are present in a document. When some hidden components change [2], the behaviour of the document gets changed.

The remaining part of this paper is as follows: Related Work is summarized in Section II. The Comparative study of various page change detection algorithm is summarized in Section III. Problem Statement is summarized in Section IV. Proposed Work is presented in Section V. Conclusion and Future Work is summarized in Section VI.

2. RELATED WORK

Number of research papers was found that handled the design of efficient algorithms for detecting changes [11] in Web pages.

H. P. Khandagale and P. P. Halkarnikar [4] produce system based on Node Signature approach relates to HTML pages. HTML webpages convert into XML webpage and then renovates XML webpages to parse trees using DOM. It compares both the trees using hash values. Hash value is generated using get hash function which produced value randomly.

Yadav [5] proposed checksum (hash value) based content level change detection. At the time of page crawling, only comparison was made to the text code of that page. The main drawback of this technique is that if any change in that

value is detected for the actual copy on the web as compared to the local copy, regardless it is significant or not, the page was refreshed or re-crawled. Hence this technique results in network overload and wasted resources for the search engine.

Vidushi and Sachin[8] assign a code based on Unicode to the text content appeared in a web page. They did the process of assigning code word by word. Whenever difference is found between two words we came to the conclusion that page has been updated otherwise comparison is made till the end of the web page.

S. Mali and B.B. Meshram [10] proposed architecture which uses Re-visit policy based approach. This architecture determines three layers: Page relevance computation, determination of page change and update the URL repository. Crawler gets the URL and parses the web page and then discovers the relevancy of web pages. After that change has detected and at last it updated the repository.

3. COMPARATIVE STUDY OF DIFFERENT ALGORITHMS FOR WEB PAGE CHANGE DETECTION

Algorithm	Speed	Space	Use	Issues	References
Level order Traversal [2007]	Faster	Medium	Reduce Network traffic, less cost, Simple	Create Insufficient Information for the user	[10]
Optimized Hungarian algorithm [2007]	Slow	Small	Good details of time and accuracy analysis, Good performance results	More Running time, Inefficient for user selected zone	[6]
Hashing Based Algorithm [2008]	Faster	Small	Less computation time, Used for RSS feeds change detection	Sometimes not support change delete operations, not good when root node is changed	[13]
Node signature comparison [2010]	Faster	Depends upon number of nodes	Suitable for attribute and text changes	No discussion about accuracy and running time	[4]
Tree Traversing [2012]	Faster	Depends upon number of nodes	Simple to understand, less browsing time	Performance not defined when depth of tree is more	[18]
Document Tree based Approach [2013]	Faster	Depends upon number of nodes	Good comparison study of different algorithms, simple to understand	Comparison is difficult if more number of nodes	[14]

document index based change detection technique[2013]	faster	medium	the pages, which are not significantly modified, are not retrieved and the pages which are significantly modified, only their document indices are retrieved.	No discussion about running time	[12]
Web page modification detection system at multiple nodes [2013]	Faster	Depends upon number of nodes	detect the structural as well as content changes at multiple nodes at one time, multiple changes are found in the web page	Comparison is difficult if more number of nodes	[15]
Web Page Change Detection System for Selected Zone [2014]	Faster	Depends upon number of nodes	Helps to locate minor or major changes within the selected zone of document. Use for stock broker, job seekers who are continuously monitoring the changes in the web Pages.	Comparison is difficult if more number of nodes	[16]
Web page change detection using document tree based method[2015]	Faster	Depends upon number of nodes	Efficient to detect the structural changes perfectly if the constructed tree represents the true Hierarchical relationship among tags.	If nesting of tag structures is misaligned then constructing the document tree becomes difficult	[17]

4. PROBLEM STATEMENT

Dynamically changing nature of web requires that mobile web crawlers must be able to intelligently decide about new pages and the changes in already crawled pages, which pages are to revisit and which pages are to skip. The ability of a mobile web crawler to decide carefully about the pages to be transferred through the network, results in significant reduction of load on the resources of the web server and underlying network. For reducing network load, an efficient technique for mobile web crawler for revisiting the already crawled pages and select only those pages which have been changed for downloading purpose is to be developed.

In the existing work proposed by Mali [20], to detect any content level change in the web page the author uses frequency of the character in a particular word for finding the text code.

The drawback of using the frequency is that if any two words having same total number of characters and the number of times any character occur in the word is also same but the words are differ in the position of the characters would give rise to the same text code i.e. text code based on frequency does not reflect the changes made in any word by changing the position of the character. e.g if the original web page contains the word "INTERNET" and the modified web page contains the word "TERNENIT".Both the words contains same characters but the positions are different, the text code of both the words calculated by method proposed by Mali[20] turns out to be same. To overcome the above mentioned problem an efficient method for finding the text code for the web page is to be developed.

5. PROPOSED WORK

To overcome the problem as stated earlier, a unique text code is assigned to each and every word appearing in the web page. The Code assigned has been called text code. Here character digit position in the word is used instead of frequency of character.

The text code based on position value of character has been assigned to initial web page and modified web page. If the text code of both the pages is same it means there is no change in the web page otherwise webpage has been changed and crawler have to download it again and refresh it. The code has been assigned word by word. The users have to input the old web page and modified web pages in which the changes are to be detected. The input module then asks the crawler to fetch the two pages. The old web page will be fetched from archive and the new web page will be fetched from the site. The comparison algorithm module find the difference between the two web pages by finding the text code assigned to each and every word. If the web

page had not been changed since last visit, then that web page will not be downloaded again and thus network load is reduced

Whenever we found the difference between two words we came to the conclusion that page has been updated otherwise comparison is made till the end of the web page.

The proposed method offers the following advantages,

- The text contents on a particular page have been assigned a unique code.
- The minute changes like changing the position of character occurring in the text also get noticed.
- The text code of two web page is compared word by word, there is no need to store the whole web page for indexing purpose
- The formula uses ASCII values as each character or symbol has unique ascii values so there is no chance for arising ambiguity.

$$\text{The formula for text code is } \text{text code} = \frac{\sum (10^{\text{position value of the character}} * \text{ASCII code of the character})}{\text{Distinct character count}}$$

The position value of character is the character digit position i.e. 0,1,2 etc.

ASCII code is the ascii value of the character.

Distinct character count is total of different characters appearing in a particular word.

6. CONCLUSION AND FUTURE WORK

Web page change detection system selects the zone provided by the user, and then the system will detect the Changes in old and new pages. The researchers had worked in structural and content changes. Some said that presentation changes are irrelevant. The proposed Algorithm works on text code based approach. The text code based on position for content change removes the ambiguity. The future work for this approach is that it will show the change detection in the images as well.

REFERENCES

- [1] Imad Khoury, Rami M. El-Mawas, Oussama El-Rawas, Elias F. Mounayar, Hassan Artail, "An Efficient Web Page Change Detection System Based on an Optimized Hungarian Algorithm", IEEE, Vol. 19, No. 5, May 2007
- [2] Vipul Sharma, Mukesh Kumar, Renu Vig, "A Hybrid Revisit Policy For Web Search", Journal of advances in information technology, vol. 3, no. 1, February 2012.
- [3] Yadav D. 2009 "Design of A Novel Incremental Parallel web crawler" Phd thesis, Jaypee Institute of Information Technology University, 2009.
- [4] H. P. Khandagale and P. P. Halkarnikar, "A Novel Approach for Web Page Change Detection System", In International Journal of Computer Theory and Engineering, Vol. 2, No. 3, 1793-8201, pp 364-368, June, 2010
- [5] Yadav D, Sharma AK, Gupta JP. Topical web crawling using weighted anchor text and web page change detection techniques. JWSEAS Trans Inform Sci Appl Arch 2009;6(2):263-75.
- [6] Khandagale HP, Halkarnikar PP. A novel approach for web page change detection system. Int J Comput Theory Eng 2010; Vol. 2, No. 3, pp 364-367, June, 2010.
- [7] Srishti Goel, Rinkle Rani Aggarwal "An Efficient Algorithm for Web Page Change Detection", International Journal of Computer Applications (0975 – 888) Volume 48– No.10, June 2012
- [8] Vidushi Singhal and Sachin Sharma "Text content based web page refresh policy" Journal of Global Research in Computer Science, Volume 3, No. 11, November 2012
- [9] S mali, B B Meshram "Implementation of multiuser personal web crawler", Software Engineering(CONSEG), CSI 6th International conference, 2012
- [10] D.Yadav, A.K.Sharma, J.P.Gupta "Change Detection In Web page" in a proceeding of 10th international conference on information technology, pp 265- 270, 2007.
- [11] S.Goel, R.R.Aggarwal "Comparative Analysis of Webpage Change Detection Algorithms" International Journal of Research in Engineering & Applied Sciences, vol2, issue 2, PP. 1382-1397, February 2012.
- [12] Badawi a, Mohamed a, Hussein b, Gheith a, "Maintaining the search engine freshness using mobile agent", Egyptian Informatics Journal ,27-36, 2013.
- [13] Hassan Artail, Kassem Fawaz, "A fast HTML web page change detection approach based on hashing and reducing the number of similarity computations", Data & Knowledge Engineering 66 (2008) 326-337
- [14] Varshney Naveen Kumar and Dilip Kumar Sharma "A Novel Architecture and Algorithm for Web Page Change Detection" IEEE IACC, 782-787, 2013.
- [15] Neha Batra, Chandna Jain "A Novel Approach on Web Page Modification Detection System at multiple nodes", International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 9, September 2013
- [16] Sandesh D. Jain, H.P.Khandagale" A Web Page Change Detection System For Selected Zone Using Tree Comparison Technique,



International Journal of Computer Applications Technology and
Research Volume 3– Issue 4, 254 - 262, 2014

- [17] Annu Saini, Taruna Kumari, "A review on Web page change
detection using document tree based method", International Journal
of Emerging Trends & Technology in Computer Science , Volume 4,
Issue 5(2), September - October 2015