IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified ⁜ Impact Factor 8.066 ⁜ Vol. 10, Issue 3, March 2023

DOI: 10.17148/IARJSET.2023.10322

Study of Web Page Change Detection to Reduce Network Load

Dr. Kompal

Govt. College, Panchkula

Abstract: With continuously growing of the Web size, it is becoming extremely difficult to search it for relevant information. Larger part of the bandwidth and internet traffic is consumed by the Web crawlers that collect pages for indexing by different search engines. Web page change detection refers to the process of monitoring a website or web page for any modifications, updates, or new content. The load on the remote server is also caused by crawlers by using its CPU cycles and memory.

I. INTRODUCTION

The mobile crawlers filter out pages that are not amended since the last crawl before sending them to the search engine for indexing purpose. Change detection methods help to identify whether web documents maintained at Search engine end have been modified at Web server end or not and if so then the modified copy should replace the existing copy at Search engine end to keep the Search engine's repository up-to-date. Tremendous amount of data sources are available online. Web surfers are flooded with huge collection of web pages managed by various sources. It is quite possible that after downloading web pages, their local copies residing in the repository/database of a search engine become obsolete compared to the copy on the web server end. Therefore, need arises to refresh the web pages of database at regular interval. The type of change on World Wide Web are:

- **Structural Changes**

Structural changes occur whenever a tag is added or deleted in the web page i.e. addition or deletion of a tag effects structural changes [1] in a web page. Sometimes the addition /deletion of a link also effect a structural change. These kinds of amendments are very important to discover as they are not visually traceable.

- **Content or Semantic Changes**

Semantic changes occur whenever the content of a web page varies according to the reader point of view [2]. A stock trader may be interested to know the updated status of the market or the latest price of the share. He is interested in the latest or the updated status of the market and not in the earlier price or the earlier market status.

- **Presentation or Cosmetic Change**

Presentation or Cosmetic types of changes occur whenever the appearance of a web page is changed but the content of a web page remains the same [3]. For example with the modifications in tags may change the presentation of a page without change in the content of a page.

- **Behavioural Changes**

Behavioural changes refer to changes in the active components which are present in a document [4]. For example, when hidden components such as scripts, applets changes, the behaviour of the document gets changed. However, it is difficult to find out such changes when the codes of such type of active components are hidden in other files.

1. Crawling Policies

The behaviour of a Web crawler is the outcome of a combination of policies [5]:
Selection policy*:* This policy states that a crawler always downloads only a portion of the Web pages, it is highly desired that the most relevant pages are downloaded and not just a random sample of the Web. Designing a good selection policy is a challenging task.

A re-visit policy*:* This states when to check for Web page changes. The Web has a very dynamic nature, and crawling a fraction of the Web can take a really long time, usually measured in weeks or months. By the time a Web crawler has finished its crawl, many events could have happened. These events can include creations, deletions and updates. From

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified ⁜ Impact Factor 8.066 ⁜ Vol. 10, Issue 3, March 2023

DOI: 10.17148/IARJSET.2023.10322

the search engine's point of view, there is a cost associated with not detecting an event, and thus having an outdated copy of a resource.

A politeness policy*:* This states how to avoid overloading Web sites. Needless to say, if a single crawler were performing multiple requests per second and/or downloading large files, a server would have a hard time keeping up with requests from multiple crawlers.

A parallelization policy: This states how to coordinate distributed Web crawlers. A parallel crawler is a crawler that runs multiple processes in parallel. The goal is to maximize the download rate while minimizing the overhead from parallelization and to avoid repeated downloads of the same page[6]. To avoid downloading the same page more than once, the crawling system requires a policy for assigning the new URLs discovered during the crawling process, as two different crawling processes can find the same URL.

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

Algorithm	Speed	Space	Use	Issues
Node signature comparison [2010]	Faster	Dependency on number of nodes	Suitable for text and attribute changes	No discussion about accuracy and running time
Tree Traversing [2012]	Faster	Depends upon number of nodes	Simple to understand, less browsing time	Performance not defined when tree depth is more
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
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
Web page change 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	More number of nodes makes comparison difficult

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified ⁜ Impact Factor 8.066 ⁜ Vol. 10, Issue 3, March 2023

DOI: 10.17148/IARJSET.2023.10322

A Web Page Change Detection System For Selected Zone Using Tree Comparison Technique[2014]	Faster	Reduce browsing time	Generalized tree comparison Algorithm for selected zone	Web Page Detection system is for selection zone using generalized Technique detects changes for selected zone
Change Detection Optimization in Frequently Changing Web Pages[2017]	Faster	medium	Automated the process of change detection	Only supports the optimization of current change detection systems

III. CONCLUSION

Due to the highly dynamic nature of the web, it is becoming extremely difficult for a search engine to provide the latest set of information to the user as the Web pages are amended very frequently. So, it is becoming a necessity to develop a Web page change detection system which provides relevant information in the least browsing time and reduce network load by downloading only the changed Web pages.

REFERENCES

[1]. Y. Wang, D. J. DeWitt and J. Y. Cai, "X-Diff: an effective change detection algorithm for XML documents," in 19th International Conference on Data Engineering, Bangalore, India, 2003, pp. 519-530.
[2]. S. Chakravarthy and S. Hara, "Automating Change Detection and Notification of Web Pages (Invited Paper)," in 17th International Workshop on Database and Expert Systems Applications, Krakow, Poland, 2006, pp. 465-469.
[3]. 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
[4]. G. Srishti, Rinkle R. A. "An efficient for web page change detection" , IJCA, VOL. 48, NO.10, June, 2012.
[5]. Shobhna, Manoj Chaudhary "A Survey on Web Page Change Detection System Using Different Approaches" International Journal of Computer Science and Mobile Computing", Vol. 2, Issue. 6, June 2013, pg.294 – 299.
[6]. S. D. Jain and H. Khandagale, "A Web Page Change Detection System For Selected Zone Using Tree Comparison Technique," International Journal of Computer Applications Technology and Research, vol. 3, no. 4, pp. 254 - 262, 2014