A Novel Technique for Spare Web Page Detection in Parallel Web Crawler

Gaurav Kumar Srivastav
Research Scholar
NIET, Greater Noida

Irphan Ali
Asst. Professor
NIET, Greater Noida

Atul Kumar Srivastava
Asst. professor
Amity University, Noida

ABSTRACT
The World Wide Web is increasing in the random rate of web pages and all web pages are rapidly updated about the need of user. Web search engine downloads web pages and the user cannot take the relevant update information for World Wide Web within short period of time. In this paper, we represent novel technique which helps in downloading the updated relevant web pages from World Wide Web. We will be implementing a new algorithm which can find out the update web page on World Wide Web. This algorithm compares the Content Weight of old web page content and downloaded update web page content.

In this paper, we have also avoid the downloading of spare web pages from World Wide Web .This is a novel techniques improved the downloading rate of web pages and it is decreased the network bandwidth of web crawler by the help of parallel web crawler. This web detection technique will be downloaded the update web pages from World Wide Web and minimize the web browsing period of time.

Keywords
World Wide Web, Search Engine, Web Crawler, Parallel Web Crawler, ASCII value and Position of i^{th} .

1. INTRODUCTION
The World Wide Web also known as "Web, or "W3". WWW is share the global of network-accessible information such as images, video, commercial and search the any information on internet, the World Wide Web is an example of human knowledge.

It easily shares the information between people. That means it allows anyone to read and publish information in the form of Web documents freely. The World Wide Web cannot show other details of the user such as communication protocols, machine locations, and operating system. It allows users can be access to any other Web pages without any restrictions. According the Brin (1998):

"***The Web is a large collection of completely uncontrolled different documents (video, image, file etc.)"***

The Web is accessible through web pages on the browser. They access the information on web by Search engine [1]. Search engine finds all the information which is requested by user for Internet. we can be seen as a large shapeless and global database by the help of World Wide Web. These pages are continuously updated and web crawlers searches and downloads pages in the repository. Web crawler has to find quality pages and updated pages into its repository. There is a need to develop a novel technique to manage, retrieve, and update information in the World Wide Web. The goal of this paper is to find solutions to downloaded updated web pages for the World Wide Web efficiently. In other words, how do we develop a fast, robust, and accurate web document by search engines and downloaded the update web document from World Wide Web.

2. RELATED WORK
S. chawathe and H. Garcia-Molina [2] uses Ladiff algorithm based on hierarchical structural information algorithm. In this algorithm compares node of the sub tree in the web pages. It has linear time complexity.

L. Francisco-Revilla, F. Shipman, R. Furuta, Unmil Karadkar, and Avital Arora [3] define the Walden's Paths Path Manager approach. This path related for web pages. It takes a path file as input web page and checks all the paths for accessing the web pages. After that, it retrieves the all web pages then parses the text content of web page. For detecting the changes, by the use of Johnson's and Proportional's Algorithm.

Y. Wang, D. DeWitt, and J. Cai [4] proposed an algorithm, which is do the three different operations: Insert, Delete and Update. In this algorithm uses use the three steps: Parsing and Hashing, Matching and Generating minimum cost edit script.

S. Chakravarthy and S. C. Hari Hara [5] proposed a system WebVigiL, which detects the change in web page and time based warning of HTML/XML pages based on user individual changes of interest. WebVigiL is used as an information monitoring and notification system, which retrieves information or data from remote servers, detects changes, and notifies the changes of users of their interest.

Ying Pan and Xuhua Ding [6] propose a new method for web detection, which is used to avoid phishing the content in the web pages. It is represent the idea of anomalies in web pages. Using this methodology, we find the text content changes in the web page and match with the previous web pages.

Imad Khoury, Rami M. El-Mawas, Oussama El-Rawas, Elias F. Mounayar, and Hassan Artail [7] presents optimized Hungarian Algorithm based approach to access user requests, to procure Web pages from the WWW, to allow users for selecting zones in Web pages to monitor, and to highlight the changes on the Web pages. The main purpose of Hungarian algorithm is to find an optimal solution of the assignment problem. In this algorithm, Authors discuss about the various type of module such as Interface Modules, Storage Modules and Algorithmic Modules.

D. Yadav, A.K.Sharma and J.P.Gupta [8] describe a method for building a system to monitor the changes to Web pages. The author uses to proposed new algorithm, which is find text content changes in the Web pages. This algorithm is use the three phases: document tree construction, document tree encoding and tree matching.

Hassan Artail and Michel Abi-Aad [9] describe an approach in which the elements of the same type in the two versions of the web page have been restricted. Before defining the similarity computations, HTML web page has been changed in to the XML page. Then it compares all nodes of the trees and then finds the similarity computations.

D. Yadav, A.K. Sharma and J.P. Gupta [10] proposed a novel architecture for produce a parallel crawler. There are two types of changes in web pages. First, Changes in Page Structure and Second, Changes in Text Content.

H.Artail and K. Fawaz [11] proposed the change detection across two versions of a page as old web page and new web page is produced by performing similarity computations after changing the page in an XML in which a node corresponds to HTML tag.

H. P. Khandagale and P. P. Halkarnikar [12] 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 gethash function which is produced value randomly.

S. Mali and B.B. Meshram [13] 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 updates the repository.

S. Goel and R. R. Aggarwal [14] use the hash based algorithm used to detect the changes. First a web page is searched in which changes will be detected. After that the tree is designed for that particular web page and after that the two trees are compared by the tag values assigned to each node. Authors used the Bottom up approach for change detection X Diff Algorithm based approach.

3. ISSUE OF CHALLENGE

In this paper we are detecting the content change in web page, if the positions of the some words are changed. It is called content change. So the challenge is that removing the ambiguity is a very typical task. So the time complexity of that approach is high. So minimizing the high complexity is difficult to solve. Accessing the accuracy of detecting the updated web pages in WWW is a very critical step. So search engine is finding more accurate and appropriate changes in web pages in very short time. It is very difficult to download the updated web pages from the World Wide Web by web crawler.

4. PROPOSED WORK

According to definition of Lee"WWW is the collection of web pages, interconnected one to another page". This web page is combination of content of web page and HTML tags. In this paper, we will propose the new web detection algorithm about the content change in web page. It is represent the algorithm for detecting content change in web pages, and calculate the content weight for equation (3) for detecting the changes in contents.

World Wide Web=Σ (Interconnected of web pages)
-Equation (1)

Web Page = Σ (Content of web pages + HTML tags)
-Equation (2)

Content Weight=Σ (Position value of i^{th} character * ASCII value of i^{th} character)
-Equation. (3)

For equation 1 and 2 define the web page changes are depending on the content and HTML tags etc.

Algorithm for finding the content change in web page:

Assume the two Input web pages: First web page (P1), and Second web page (P2).

1. Firstly, Initialize the position value of i^{th} character of each page and calculate the Content Weight.
2. Find the length of Content (L) = strlen(word).
3. Set i=1, and repeat the step 4 and 5 until word [i^{th}]='\0'.
4. Set int ascii value (A) =int (word [i^{th}]).
5. Set P=P+(i^{th} *A)
6. Then, Calculate Content Weight=Sqrt(P)
7. Compare the Content Weight among the web pages
   If (Content Weight (P1) =Content Weight (P2)) then
   return ("Not download the web page")
   Else
   return ("download the web pages")

The Architecture of web page changes detection system is shown in Fig. 1. Web Crawler selects all the URLs which are given by Search engine. After that, it takes response from the web server about the web pages that this web page is existing

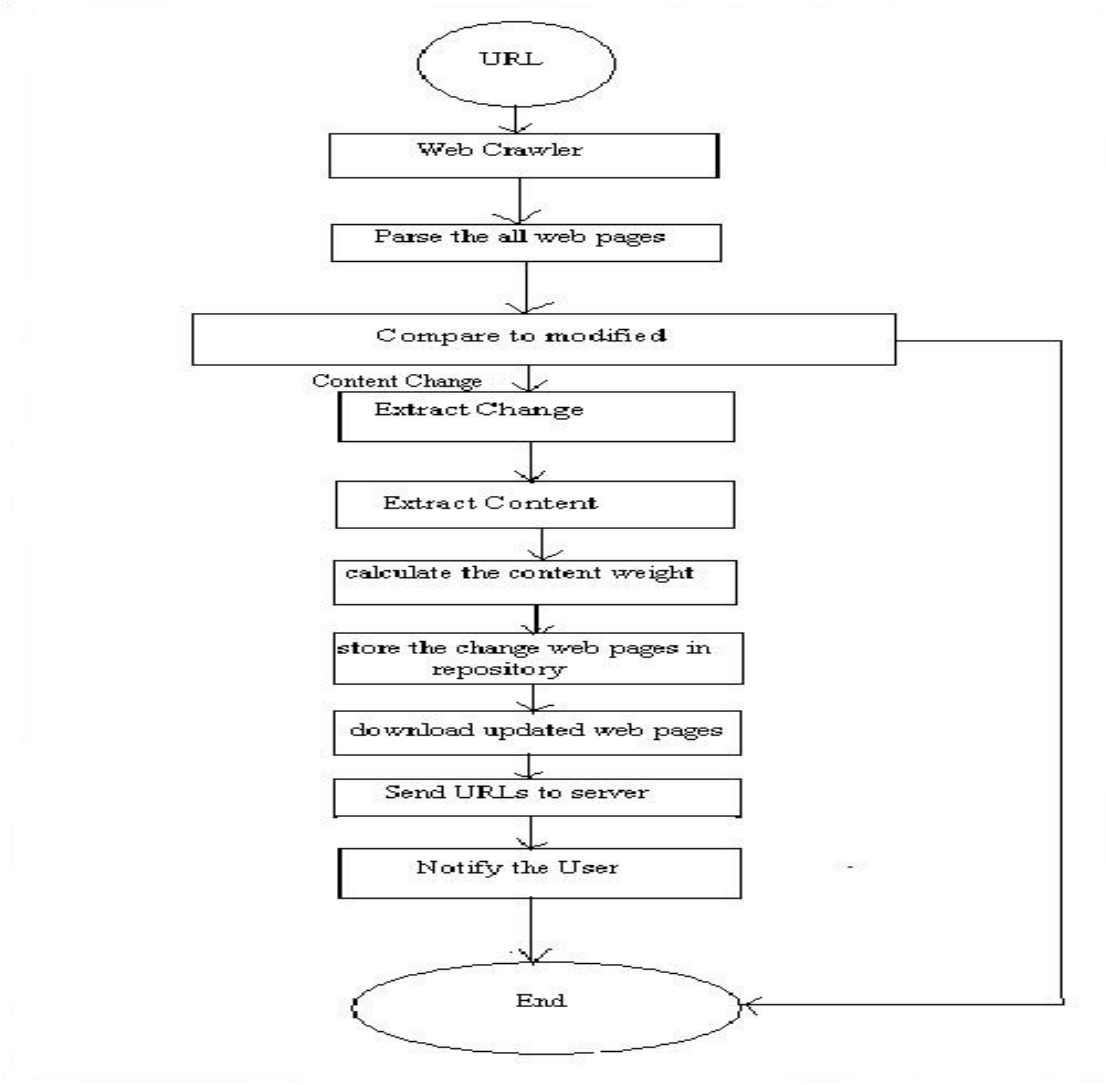

Fig (1) Architecture of web page change detection System

on the web or not. Then it parses the web pages and calculates the Content Weight for content changes. After detecting the changes, the page is downloaded. If its URL does not exist on the web then it is discarded. Then this page is sent to the server and it sends the notification to the user about the changes.

The working of architecture is represented in step by step is given below:

Step 1: Crawler downloads the Urls of the web pages.
Step 2: The page is parsed all web pages.
Step 3: The initial page is compared to the modified page.
Step 4: Extract the changes. If there is content change, calculate the Content Weight for it.
Step 5: After that these changes are stored in the repository for the future use.
Step 6: Download the update page which has occurred changes.
Step 7: Send this downloaded page to the Server.
Step 8: At last, Server notifies that changes to the user.

5. RESULT

In this portion, we are analysis the result about the proposed web detection algorithm, and also compare the content weight value of initial content, and modified content of web pages. we are taking the some example for calculate the content weight value, then compare between them.

Using formula of Content Weight in equation (1), it firstly finds the position value of the character and multiplied by its ASCII value of that particular character and stores the value in variable, which is denoted by. After that we are find the content weight value by the Square root value P.

Example 1.

http://www.landsofmaharashtra.com/agriculturalland.html

The Initial content is:

"agriculture" includes horticulture, poultry farming, the rising of crops, fruits, vegetables, flowers, grass or trees of any kind, breeding of livestock including cattle, horses, donkeys, mules, pigs, breeding of fish and keeping of bees, the use of land for grazing, cattle and for any purpose which is ancillary to its cultivation or other agricultural purpose. **Classification of agricultural lands as per its use for cultivation.**

The sum of ASCII characters with position value: 8861186

Total Character count: 433
Then Content Weight: 2976.774429

The modified page content is:

"agriculture" includes horticulture, poultry farming, the rising of crops, fruits, vegetables, flowers, grass or trees of any kind, breeding of livestock including cattle, horses, donkeys, mules, pigs, breeding of fish and keeping of bees, the use of land for grazing, cattle and for any purpose which is ancillary to its cultivation or other agricultural purpose.

The sum of ASCII characters with position value: 6241634
Total Character count: 365
Then Content Weight: 2498.326240

Example 2.
http://www.hcltech.com/manufacturing

The Initial content is:

In today's dynamic and uncertain global marketplace, manufacturing remains an important driver of innovation. As manufacturers internationalize their footprints, visibility and collaboration across supply chains are increasingly critical. And as the sector faces new technologies, products, and ways of working, manufacturers need new strategies to stay ahead of the competition.

The sum of ASCII characters with position value: 6894295
Total Character count: 379
Then Content Weight: 2625.681816

The modified page content is:

In today's dynamic and uncertain global marketplace, manufacturing remains an important driver of innovation. As manufacturers internationalize their footprints, visibility and collaboration across supply chains are increasingly critical. And as the sector faces new technologies, products, and ways of working, manufacturers need new strategies to stay ahead of the competition.
These are provide the following services such as Sourcing, Product engineering and manufacturing, design and development, Supply chain management, After-market services, Logistics and distribution, Sales and marketing, Service frameworks.

The sum of ASCII characters with position value: 18201974
Total length of Character count: 618
Then Content Weight: 4266.377152

Example 3.
http://www.birlasoft.com/SERVICES/ENTERPRISEAPPLICATIONS/PeopleSoft.aspx

The Initial content is:

Enhancing productivity and business performance, while developing better customer relationships and keeping total cost of ownership at a minimum, can be a challenge without the right tools and services. Birlasoft provides comprehensive PeopleSoft solutions, which allow organizations to address all their business needs with integrated solutions and custom features. As a Global Oracle Platinum partner, our dedicated PeopleSoft Center of Excellence (CoE) enables us to deliver end-to-end consulting services that are tailor-made to suit all client requirements, **ensuring they get** maximum value from their PeopleSoft investment. Leveraging our significant Oracle knowledge, Birlasoft's Oracle **CoE** has been helping clients replace older legacy systems, integrate 3rd party bolt on systems, improve business results and consequently increase return on investment.

The sum of ASCII characters with position value: 35367648
Total length of Character count: 861
Then Content Weight: 5947.070539
The modified page content is:

Birlasoft provides comprehensive PeopleSoft solutions, which allow organizations to address all their business needs with integrated solutions and custom features. As a Global Oracle Platinum partner, our dedicated PeopleSoft Center of Excellence (CoE) enables us to deliver end-to-end consulting services that are tailor-made to suit all client requirements. it is maximum the value from their PeopleSoft investment. Leveraging our significant Oracle knowledge, Birlasoft's Oracle Center of Excellence has been helping clients replace older legacy systems, integrate 3rd party bolt on systems, improve business results and consequently increase return on investment.

The sum of ASCII characters with position value: 21223858
Total length of Character count: 667
Then Content Weight: 4606.935858

So, we have taken the two content weight value of initial content and modified content page. If the content weight value of initial content is equal to the content weight value of modified content. Then, we can say that modified content is equal to initial content, there are no changes in text content among them. Otherwise the text is not same between them.

6. CONCLUSION AND FUTURE WORK

Propose web detection algorithm is used for the text content detection, these algorithm detect the changes in text content in the web page. This text content is provided by the user by updating of web page, and after that the system will detect the content changes in web pages and compare the Content Weight of old web page to updated web page. If the Content Weight values are different, then crawler download the update web pages otherwise discard the downloading. In this paper, researcher had worked on content changes and proposed the new Algorithm for content change in web pages. The propose web detection algorithm is not evaluate the 100% accuracy about the result. So, the future work for user that improve the performance of propose web detection algorithm, and as well as do in future work about the change detection in the images, audio, and video in web pages

7. REFERENCES

[1] Sergey Brin and Lawrence Page, "The Anatomy of a Large-Scale Hypertextual Web Search Engine", In Proceedings of the Seventh World-Wide Web Conference, 1998.

[2] S. Chawathe, H. Garcia-Molina, "Meaningful Change detection in structured data", In proceeding in ACM SIGMOD International conference,pp 26-37,May1997.

[3] L. Francisco-Revilla, F. Shipman, R. Furuta, Unmil Karadkar, and Avital Arora," Managing Change on the Web", ACM /1-58113-345- 6/01/0006 pp 67-76, June 2001.

[4] Y. Wang, D. DeWitt, and J. Cai, "X-Diff: An Effective Change Detection Algorithm for XML Documents", Proc. 19th Int'l Conf. Data Eng., pp. 519-30, 2003.

[5] S. Chakravarthy and S. C. Hari Hara, "Automating Change detection and Notification of Web Pages", In Proceedings of the 17th International Conference on Database and Expert Systems Applications (DEXA'06), IEEE , 0-7695-2641-1/06, 2006.

[6] Ying Pan, Xuhua Ding "Anomaly Based Web Phishing Page Detection",In Proceedings of the 22nd Annual Computer Security Applications Conference,IEEE, 0-7695-2716-7/06,2006.

[7] Imad Khoury, Rami M. El-Mawas, Oussama El-Rawas, Elias F. Mounayar, and Hassan Artail, "An Efficient Web Page Change Detection System Based on an Optimized Hungarian Algorithm ", In IEEE Transactions on Knowledge and Data Engineering, Vol. 19, NO. 5, pp 599-613, May 2007.

[8] D. Yadav , A.K.Sharma and J.P.Gupta,"Change Detection in Web Pages ", In 10th International Conference on Information Technology ,IEEE , 0-7695-3068-0/07, pp 265-270, 2007.

[9] H. Artail and M. Abi-Aad,"An enhanced web page change detection approach based on limiting similarity computations to elements of same type", Springer Science + Business Media, LLC, pp 1-21, 2007.

[10] D. Yadav , A.K. Sharma and J.P. Gupta, "Parallel Crawler Architecture and Web Page Change Detection ", In WSEAS Transactions on Computers, ISSN: 1109-2750 , Issue 7, Vol. 7, pp 929-940 , July 2008.

[11] H.Artail , K. Fawaz. ,"A fast HTML web page change detection approach based on hashing and reducing the number of similarity computations", Elsevier, Data & Knowledge Engineering 66 , pp 326– 337, 2008.

[12] 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 .

[13] S. Mali and B.B. Meshram , "Focused Web Crawler with Page Change Detection Policy ", In International Journal of Computer Applications (IJCA), pp 51-57, 2011.

[14] S. Goel and R. R. Aggarwal, "An Efficient Algorithm for Web Page Change Detection" , In International Journal of Computer Applications (0975 – 888), Vol. 48– No.10, pp 28-33, June 2012.