WEB PAGE CHANGE DETECTION THROUGH MACHINE LEARNING ALGORITHMS

R.Kousalya MCA., M.Phil. (Ph.D)
Head Department of Computer Applications
Dr.N.G.P Arts and Science College, Coimbatore.
kousalyacbe@gmail.com

J.Rubana Priyanga M.Phil Scholar
Dr.N.G.P Arts and Science College, Coimbatore.
rubanapriyangacbe@gmail.com

ABSTRACT

This research work is mainly used to detect changes in the web page. The existing Case per technique is taken for research and compared with the proposed technique Arti-Q Dynamic web page change detection. Artigence technique is an Arti-Q algorithm, used for reducing response time for each request. The Arti-Q technique introduces two heuristics method such as internal proxy and external proxy. The methodology used for web page change detection are Net Tracker which is used to track the commodity market server to detect the commodity rate changes in Gold, silver, INR, Palladium, Platinum, Rhodium and Crude. The proposed Arti-Q DWPD technique receives the URL as a parameter to detect the HTML source code and using substring comparison to filter out the commodity rate changes. The detected rates are compared with the lastly retrieved rate and if there is any change identified between them, it will be displayed in the list view control and for every change detection, the varying count gets increment the value by 1. The process repeated until the threshold time exceed as for every 0.5 seconds. The experimental results shows that the implementation improves an accuracy level while compared with existing implementation.

Keywords: Case per technique, Arti-Q DWPD technique, Net Tracker, internal proxy and external proxy.

INTRODUCTION

Web mining is a type of data mining is used to extract web data from web pages. As data mining basically deals with the structured form of data, web mining deals with the unstructured and semi structured form of data. Web mining consist of three techniques such as web content mining, web structure mining and web usage mining for web data extraction. The World Wide Web has evolved from a simple and static content source into a richly dynamic information deliver channel. Hence, tracking the changes in web pages has become an interesting research problem would solve the problem and facilitate users to efficiently utilize the information that is provided by web pages.
WEB USAGE MINING IN DYNAMIC WEB PAGE CHANGE DETECTION

The dynamic web page change detection is processed under the web mining concepts, and the web usage mining is the basic component being used to detect changes in web page. Web usage mining is the application of data mining techniques to discover interesting usage patterns from web usage data, in order to understand and better serve the needs of web-based applications. Web usage mining is a process of mining useful information from server logs. Usage data captures the identity or origin of web users along with their browsing behavior at a web site.

Web usage mining can be classified further depending on the kind of usage data considered as web server data, application server data and application level data. In Web Server Data, user logs are collected by the web server and typically include IP address, page reference and access time. In Application Server Data, commercial application servers such as Web logic, Story Server, have significant features to enable E-commerce applications to be built on top of them with little effort. A key feature is the ability to track various kinds of business events and log them in application server logs. In Application Level Data, new kinds of events can be defined in an application, and logging can be turned on for them, generating histories of these events. It must be noted, however, that many end applications require a combination of one or more of the techniques can be applied.

CHANGE DETECTION AND NOTIFICATION

Change detection and notification (CDN) refers to automatic detection of changes made to World Wide Web pages and notification to interested users by email or other means. Whereas search engines are designed to find web pages, CDN systems are designed to monitor changes to web pages. Before change detection and notification, it is necessary for users to manually check for web page changes, either by revisiting web sites or periodically searching again. Efficient and effective change detection and notification is hampered by the fact that most servers do not accurately track content changes through Last-Modified headers.
**PROBLEM DEFINITION**

Due to limited network and computational resources, it is often difficult to monitor the sources constantly to check for changes and to download changed data items to the copies. Finding the changes between different versions of the web pages is the core operation of web page change detection system. Users who visit a web page repeatedly at frequent intervals are more interested in knowing the recent changes that have occurred on the page than the entire contents of the web page. Because of the increased dynamism of web pages, it would be difficult for the user to identify the changes manually. The main objective is to reduce the complexity of the change-detection by focusing only on the web page change identification in which the changes have occurred.

**RELATED WORK**

In this paper, L.Liu, C.Pu and W. Tang, offers personalized delivery of change notifications as well as summarization and prioritization of Web page changes. One drawback of Web CQ is that it detects changes only between the last two versions of a HTML page. This algorithm is efficient in speed and memory space. It uses operations such as change node, delete node and insert node. Delta is constructed to find the matching of nodes between two trees. The use of XML specificities in algorithm leads to significant improvements. Drawback of this algorithm is that there is some loss of quality. Another drawback is that there is need of gathering more statistics about the size of deltas and in particular for real web data.

In this paper, Varshney Naveen Kumar, Dilip Kumar Sharma proposed a Tree based approach is good for comparing the nodes of both the tree as old web page tree and modified web page tree. It gives the relevancy to the web pages and notifies the user about detecting the changes. For detecting structural change document tree is constructed and then signature values assigned to the root nodes and child nodes of the old and new web pages are compared. This algorithm defines the good comparison study for the different algorithms and provides simple method for detecting changes.

In this paper, K.S. Kuppusamy, G.Aghila, proposed a CaSePer Model which is Change detection based on Segmentation with Personalization an enhanced model for detecting changes in the pages. The change detection is micro-managed by introducing web page segmentation. The web page change detection process is made efficient by having it perform a dual-step process. Changes in web pages can either be content changes or structural changes. The structural changes are primarily changes that occur at the template level. The content changes include modification, insertion or deletion of hypertext elements. Users navigate to web pages that they are interested in. A user interested in a specific topic might visit web pages that are related to that topic at regular intervals over time. Such users might be interested in knowing the recent changes that have occurred in that web page rather than seeing the entire web page.
SYSTEM ARCHITECTURE

The proposed work includes processing functionalities and interfaces for processing user request, fetching web pages from the internet allowing users to select zone in web pages to monitor. This method performs well and is able to detect the structural as well as content level changes even at the minute level and helps to locate minor or major changes within the selected zone of document. The properties of HTML page to extract the HTML source using http web request function which passes the website address as parameter where the function gets the Uniform Resource Locator (URL) of the internet resource that actually responds to the request. A web crawler could be a system for the bulk downloading of websites. A crawler starts off by placing an initial set of URLs, in a queue, where all URLs to be retrieved are kept and prioritized. The crawler gets a URL in some order from this queue, downloads the page, extracts any URLs within the downloaded page, and then in the queue it puts the new URLs. Artigence is the technique which uses the concept called Arti-Q DWPD method to reduce the waiting time request of web pages and calculate the threshold value.

![Web Page Change Detection System Diagram](image)

**ARTI-Q ALGORITHM**

**Step 1:** Create a two 2-dimensional array and named it as External proxy Ep(i, j) and internal proxy Ip(i,j), where ‘i’ represents Queue Number and ‘j’ represents substring data which holds silver, platinum, palladium, crude, rhodium and INR value.

**Step 2:** Create an Arti-Q Timer Control ‘T’ in Ip (i, j) with interval set to 0.5 seconds.

**Step 3:** Initialize predefined syntax Ps, variant V =0
Step 4: For every ‘T’ hits 0.5 second
Read Ep (i, j) extracted from the web page
Compare Ip (i, j) = Ep(i,j), if the variant detects, then increment V = V +1
Ep (i, j) =0 (Reset External proxy counter to zero).
Step 5: repeat Step 4.

RESULT ANALYSIS

The proposed methodologies on web page change detection have been effectively used in the Arti-Q Dynamic Web Page Change Detection with effective techniques. The output for the given query is processed and implemented in WEKA TOOL. The page source is extracted using the DWPD method. The URL is passed as a parameter in the direct cast method and using GET function, the page source is extracted and the value is allowed to store in a temporary array. A timer control is implemented where the page source is extracted for every 30 seconds and compare with the temporary array whether any changes in the content is made or not. If any change detected, then the changed value has been displayed to the user. The result shows the given output from various processes is better than existing process. The result has been analyzed.

<table>
<thead>
<tr>
<th>Arti-Q-DWPD</th>
<th>Duration</th>
<th>No. of Variants</th>
<th>Duplicates Detected</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1-30 minutes</td>
<td>62</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>31-60 minutes</td>
<td>57</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>61-90 minutes</td>
<td>58</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>91-120 minutes</td>
<td>49</td>
<td>0</td>
</tr>
</tbody>
</table>

(Y-axis no.of variants-axis time seconds)

Evaluated Arti-Q DWPD
From the above Arti-Q DWPD table, it is clearly stated in the first 1-30 minutes 62 variants detected in the page change detection with no duplicates identified; similarly in the next 31-60 minutes, 57 variants detected with no duplicates identified. In the next 61-90 minutes, 79 variants detected with no duplicates identified and in the last 91-120 minutes, 65 variants detected with no duplicates identified.

The experimental result shows that our proposed approach provides higher performance rate while compare with conventional methods. The threshold values exceed for every 0.5 seconds and it list the value of data rate changes for every 0.5 seconds.
CONCLUSION AND FUTURE WORK

This research work proposed Arti-Q based Dynamic web page change detection for selected zone helps us to reduce browsing time. The Web Page Detection system is for selection zone using generalized Technique detects changes for selected zone. It detects changes for text and attributes at minute level. It shows the output of change text in red colour. This saves the time of detecting the changes in the entire Web page. From the analysis of various research papers it is concluded that using a web page change detection system saves the user browsing time. Every algorithm plays some role in improving the performance of previous algorithm. Detection of changes in structure, content, presentation, behaviour of the web page helps the user to know about the updated information. The algorithm for checking page update parameters is designed to show even the smallest of change in a web page and the values stored for the parameters are distinct for small details as well. As a future work various performance parameters like running time, complexity of computation, number of nodes, depth of tree, load balancing of user requests for notification in various algorithms.
can be measured and improved. Algorithms can also be implemented in dynamic web pages or in password protected web sites.

REFERENCES


R.Kousalya M.C.A M.Phil (Ph.D), Assistant Professor, Head Department of Computer Application, Dr.N.G.P Arts and Science College Coimbatore, India. E-mail id: kousalyacbe@gmail.com

J.Rubana Priyanga M.phil Research Sholar, Dr.N.G.P Arts and Science College Coimbatore, India. E-mail id: rubanapriyangacbe@gmail.com.