A SURVEY ON WEB PAGE CHANGE DETECTION

VISHNU GOEL ,VINAY ARORA

Abstract: World Wide Web consists of billions of web pages and they are increasing at a phenomenal rate. To manage the changes, Change Detection is used which identify the differences between the two versions of the web page. It is a technique by which search engine keeps updating the current version of the web pages. In this paper, basic terms and techniques related to change detection along with its properties are presented. This literature analysis presents an overview of change detection, including the various general approaches and tools used to detect changes in the web pages and segmentation is much better approach in narrowing the concern, to the relevant portions of the web page.

Keywords: Change Detection, World Wide Web, Crawler, Continual Query.

Introduction: In the present era, where there is continuous need of updated information which is relevant for a specific period, changes have to be monitored and reported to the user as soon as possible. In monitoring, change detection system plays a significant role by reducing the human effort to minimum. To get updated information crawlers need to crawl same page after a specified interval that leads to high resource utilization and user has to frequently poll the websites of interest to get newly updated information. For this different approaches have been used for estimating the change frequency. According to recent studies 50% web pages remain unchanged during the entire period of time. And the changed 50% web pages, content change is very less. In the study, only 5% web pages changed very much from their initial version.

The experimental study performed a large crawl by downloading 151 million HTML pages and computes the checksum and figures out the rate of change. The degree of change has been classified into different groups such as complete change, large change, medium change, small change, no change. It finds out those 76% web pages fall in the group of no change and large changed web pages are 3% only and proved that incremental method is efficient in updating web indexes [1].

Change Detection: The change detection system has been used to identify structural and semantic changes in the web documents. Changes are broadly classified into four categories such as content changes, presentation changes, structural changes and behavioral changes.

A. Classification of Changes

1. Content/ Semantic changes: It refers to changes of the page contents from the users prospective. For example, a web page created for a weather forecast might be continuously updated on the hourly basis. According to weather respective to city, the page might changes in its presentation to reflect the updates more easily.
2. Presentation/ Cosmetic changes: These are changes related to the document representation that do not reflect changes in the topic presented in the document. For instance, changes to HTML tags can modify the appearance of a Web page while it otherwise remains the same.
3. Structural changes: It refers to the underlying connection of the document to other documents. For example, there are changes in the link destinations of a web directory website. The respective page might remain the same, but the fact respective links has been changed might be relevant, even if the text of the links has not. These changes are very important to detect, as they often might not be visually perceptible.
4. Behavioral changes: It refers to changes in the active components of a document. For Web pages, this includes scripts, plug-ins and applets. The effects of these types of changes are difficult to predict, as many web pages hide the script code in other files.

Related Work:

A. Using Databases

In Databases, a continual query plays a significant role in the detection of changes. A continual query consist of a triple (Q, Tcq, Stop), where Q is normal Query (e.g. written in SQL), Tcq a trigger condition, and a termination condition Stop. Continual queries can be useful both to external applications and as a convenient mechanism for implementing push-based data delivery functions beyond conventional storage, retrieval, and update of data in conventional DBMSs [2].

Ling et al. [3] introduced CQ, a project which uses continual queries which are queries written once and executed over certain data after certain interval of time to find the changes in the data. If the particular threshold has reached it informs the user about the change. It is a personalized change finder which is based on continual query concept. It combined the push and pull approach and provided the integrated environment to user. Calton et al. [2] has introduced a distributed event driven continual query system i.e. OpenCQ which includes client, server and wrapper. This system capable of push based content sensitive data management change detection. It also included various continual queries indexing techniques to improve performance and scalability.

Wei et al. [4] has provided Conquer a continual query system which is used for detecting changes over the web document. This system is capable of handling large number of continual queries by classification of them by means of their trigger structure. It also has a well defined continual query specification language.

Gedik et al. [5] has purposed a decentralized self configurable information monitoring system PeerCQ based on Continual query. A effective service partitioning scheme is introduced which integrates heterogeneity of peers with monitoring just like napster.Erwin et al. [6] motivated the use of relational database for detecting content changes in the ordered XML document. It enables the change detection for a large document by means of databases. To implement the algorithm, it used SUSXENT schema which is efficient than Xparent schema and applied various queries on leaf nodes to detect the changes.Sandeep et al. [7] has purposed a new monitoring algorithm WIC which proved to be better approach from other CQ system as it is highly parameterized and there is balance between timeliness and completeness. It is efficient and can be used in conjunction with other CQ system. The main focus of algorithm is scheduled pulling of web pages.

B. Using Diff Algorithm

From the beginning of change detection, a lot of Diff algorithms are being introduced to find the changes in the respective documents. Initially it is implemented on the document level such as LaDiff which is used for detecting the changes in two latex documents. The most common Diff is GNU Diff which uses LCS for detecting changes in two text files.

Fred et al. [8], [10] has purposed a new system WebGUIDE (Web Graphical User Interface to a Difference Engine) which is combination of two existing tools AIDE, a tool for tracking changes on the web pages and Ciao, a graphical navigator. This system enables change detection textually and graphically. It also focused on AT&T Internet Difference Engine which provides customized view of web pages for detection of changes. To find different structural changes, it uses HtmlDiff along with crawler.

Chawathe et al. [9] purposed a heuristic change detection algorithm MH-Diff i.e. meaningful Diff. This paper has provided higher quality edit script by implementing a large set of operations. The focus of algorithm is structured data which transforms the problem into edge cover that matches the nodes.

Cobena et al. [11] has purposed a more efficient and less complex diff approach for xml data warehousing for dynamic as well as static web pages. The purposed approach has focused on more accuracy rather than quality & runs in linear time in average cases. It has also tested this approach with syntactic data and found the results close to optimal solution.

Wang et al. [12] has considered an unordered tree model to be better approach for X-Diff algorithm. It has used node signature and Xhash to find the substantial difference between two Xml documents. It improved X-Diff algorithm and still provided near optimal result by means of running time. It has experienced improved X-Diff, X-Diff+ and XyDiff to prove the accuracy of result by improved X-Diff.

Anoop et al. [13] has specified that WebVigiL developed at UT Arlington, a content monitoring system that has capability of detecting changes in text as well as images but limited in web-based System. This paper focused on extending this system in the context of web by adapting and extending active technology. Chakravarthy et al. [17] has been focused on the WebVigiL change detection tool with its specification language and ECA (Event Condition and Action) Paradigm. It proposed a learning algorithm which adapts to actual page changes. WebVigiL is a modular system which is customizable according to structure of documents.

Leonardi et al. [14] has purposed a novel algorithm DTD-Diff to detect structural and semantic changes in the DTD's. The purposed algorithm has improved results over existing algorithm in case of complex DTD's and found to be less expensive as there is no conversion of DTD to XML Schema for change detection and also able to generate optimal or near optimal edit scripts or deltas.

Teevan et al. [15] has specified DiffIE an Internet explorer plugin which can be used for finding the changes while browsing the internet. For finding the corresponding changes it caches the page and compares it with previous copy of cache and highlights the differences. It improves the browsing experience by highlighting the changes.

Saad et al. [16] has specified a new approach for web archiving which uses 3 concepts i.e. the visual page segmentation, the change detection and importance of blocks in web pages. For change detection, it introduced Vi-Diff algorithm which is much better in extracting changes in visual layout structure of the document. The time of segmentation is much higher than the change detection time.

C. Using Crawler

Boyapati et al. [18] has been specified that the change detector, a site level web monitoring tool that makes use of machine learning techniques, classification and entity based components for extracting relevant changes from the web pages. The GUI of the tool provides flexibility by means of reports to subscriber.Kim et al. [19] has focused on dynamic web crawler implementation in java with the help of tracing hyperlinks in the web pages. It also figure out a new scheduling technique based on current collection cycle time(Ps), Average Cycle time(avg(Ps)) and previous collection cycle(Ps-1).It showed 59% performance benefit compared to static crawling method.

Sisi et al. [20] has specified that Surfing notes, a online tool which enables the user to annotate and archive the webpage for personal use. This tool focuses on LCS (Longest Common Subsequence) problem for detecting the text changes and possible solutions are Dynamic programming method or dominant matches by row-by-row processing technique. It also performed an experimental analysis on efficient change detection scheduling.

Rawat et al. [21] has concerned about focused crawling that uses link relevancy (TF-IDF Ranking) criteria for finding the appropriate document. To find out valuable link, it uses best first search which proves to less resource consumptive as compare to generic & search engine crawler.

D. Using Segmentation

Chang et al. [22] has specified Graphical change detection. For detecting changes, the whole document is converted into ordered labeled tree structure based on its markup language and then it is compared documents with tree matching techniques to find the edit script. It can detect paragraph changes which makes it better approach than Diff algorithm.

Seung et al. [23] has purposed a heuristic semantic change detection (SCD) algorithm for detecting semantic changes in hierarchical structured data. The considered approach does not require any preprocessing or internal structure of the source document which makes it compute frequent changes in any HTML document. The time complexity of SCD algorithm is $O((|X|x|Y|) \log(|X|x|Y|))$ where $|X|$ and $|Y|$ is number of branches in syntactic hierarchies in given documents.

Yadav et al. [24] has focused on change detection by HTTP meta data and more efficient change detection algorithm which takes linear time as compared to previous algorithms. It has been using tree encoding and level by level tree matching for structural changes. It has implemented algorithm by a specialized set of arrays to represent the relationship between nodes of tree in java.

Li et al. [25] purposed a 2 phase algorithm to detect navigation changes in webpage. It finds out block-text segments from a web page. The block text helped in deciding that whether it a navigation page or not. It evaluated DOM Based Identification with Anchor/TD and OL/PS and found to be much more efficient. It misses those pages which have one or two small Text blocks and misinterpret the pages.

Law et al. [26] has presented a hybrid web page comparison framework for detecting structural as well as visual changes. For change detection only specified region of page is considered rather than whole page. It also concluded that most relevant changes occur in visible part of web pages without scrolling.

Kuppusamy et al. [27] purposed CaSePer, a personalized change detection model which uses hybrid page segmentation method i.e. DOM tree with densitometry and introduced a "node boundary and cascaded node sequence" based segmentation technique. It has also used MD5 hashing technique to calculate signature and find changes in segments.

Varshney et al. [28] has provided the detailed and comparative analysis of different algorithm for content and structural change. It proposed an efficient algorithm for web page change detection by means of signature and text code. For algorithm, it used document tree based approach.

Conclusion: From the paper, Segmentation is considered to be most emerging and less complex change detection method and it will reduce the overall problem space and figure out changes in a more efficient way irrespective of previous approaches such as Diff algorithms and Continual Query. It helps in delivering changes to user in a more efficient way which is achieved in CaSePer.

Table I: Different Tools based on Change Detection

Criteria	**Tool Details**		
	Tools	Type Of Change	Concept Or Algorithm
Databases	OpenCQ[2]	Content	Continual Query
	WebCQ[4]	Content	Continual Query
	PeerCQ[5]	Content	Continual Query
Diff Algorithm	WebGuide[8]	Structural	HtmlDiff
	WebVigiL[13]	Structural and Content	CH-Diff and CX-Diff
	DiffIE[15]	Content	Cache
Crawler	Change Detector[18]	Structural and Content	Entity based
	Surfing Notes[20]	Content	Web Page Scheduling
Segmentation	CaSePer[27]	Structural and Content	DOM Based Segmentation

References:

1. Fetterly, Dennis, Mark Manasse, Marc Najork, and Janet Wiener. "A large-scale study of the evolution of web pages" In Proceedings of the 12th international conference on World Wide Web, ACM, 2003, pp. 669-678
2. Liu, Ling, CaltonPu, and Wei Tang. "Continual queries for internet scale event-driven information delivery." Knowledge and Data Engineering, IEEE Transactions on 11, no. 4, 1999, pp. 610-628.
3. Liu, Ling, CaltonPu, Wei Tang, David Buttler, John Biggs, Tong Zhou, Paul Benninghoff, Wei Han, and Fenghua Yu. "CQ: a personalized update monitoring toolkit." In ACM SIGMOD Record, vol. 27, no. 2, 1998, pp. 547-549.
4. Liu, Ling, CaltonPu, Wei Tang, and Wei Han. "CONQUER: A continual query system for update monitoring in the WWW." Computer Systems Science and Engineering 14, no. 2, 1999, pp. 99-112.
5. Gedik, B.; Ling Liu, "PeerCQ: a decentralized and self-configuring peer-to-peer information monitoring system," Distributed Computing Systems, 2003. Proceedings. 23rd International Conference, 19-22 May 2003, pp.490-499.
6. Leonardi, Erwin, et al. "Detecting content changes on ordered XML documents using relational databases." Database and Expert Systems Applications. Springer Berlin Heidelberg, 2004.
7. Pandey, Sandeep, KedarDhamdhere, and Christopher Olston. "WIC: A general-purpose algorithm for monitoring web information sources." In Proceedings of the Thirtieth international conference on Very large data bases-Volume 30, VLDB Endowment, 2004, pp. 360-371.
8. Douglis, Fred, Thomas Ball, Yih-Farn Chen, and EleftheriosKoutsofios. "WebGUIDE: Querying and navigating changes in web repositories." Computer Networks and ISDN Systems 28, no. 7, 1996,pp 1335-1344.
9. Chawathe, Sudarshan S., and Hector Garcia-Molina. "Meaningful change detection in structured data." ACM SIGMOD Record. Vol. 26. No. 2. ACM, 1997.
10. Douglis, Fred, Thomas Ball, Yih-Farn Chen, and EleftheriosKoutsofios. "The AT&T Internet Difference Engine: Tracking and viewing changes on the web."World Wide Web 1, no. 1, 1998, pp. 27-44.
11. Cobena, Gregory, Serge Abiteboul, and Amelie Marian. "Detecting changes in XML documents." In Data Engineering, 2002. Proceedings. 18th International Conference on IEEE, 2002, pp. 41-52.
12. Wang, Y.; DeWitt, D.J.; Cai, J.-Y., "X-Diff: an effective change detection algorithm for XML documents," Data Engineering, 2003. Proceedings. 19th International Conference, 5-8 March 2003, pp.519-530.
13. SankaAnoop, ShravanChamakura, and Sharma Chakravarthy. "A dataflow approach to efficient change detection of HTML/XML documents in WebVigiL."Computer Networks 50, no. 10, 2006, pp. 1547-1563.
14. Leonardi, Erwin, Tran T. Hoai, Sourav S. Bhowmick, and Sanjay Madria. "DTD-Diff: A change detection algorithm for DTDs." Data & Knowledge Engineering61, no. 2, 2007, pp. 384-402.
15. Teevan, Jaime, Susan T. Dumais, Daniel J. Liebling, and Richard L. Hughes. "Changing how

people view changes on the web." In Proceedings of the 22nd annual ACM symposium on User interface software and technology, ACM, 2009, pp. 237-246.

16. Saad, Myriam Ben, and StéphaneGançarski. "Using visual pages analysis for optimizing web archiving." In Proceedings of the EDBT/ICDT Workshops, ACM, 2010, p. 43.
17. Chakravarthy, Sharma, AnoopSanka, Jyoti Jacob, and Naveen Pandrangi. "A learning-based approach for fetching pages in webvigil." In Proceedings of the 2004 ACM symposium on Applied computing, ACM, 2004, pp. 1725-1731.
18. Boyapati, Vijay, Kristie Chevrier, AviFinkel, Natalie Glance, Tom Pierce, Robert Stockton, and Chip Whitmer. "ChangeDetector™: a site-level monitoring tool for the WWW." In Proceedings of the 11th international conference on World Wide Web, ACM, 2002, pp. 570-579.
19. Kim, K. S.; Kim, K. Y.; Lee, K.H.; Kim, T. K.; Cho, W. S., "Design and implementation of web crawler based on dynamic web collection cycle," Information Networking (ICOIN), 2012 International Conference, 1-3 Feb. 2012, pp.562-566.
20. Sisi He; Chan, E., "Surfing Notes: An Integrated Web Annotation and Archiving Tool," Web Intelligence and Intelligent Agent Technology (WI-IAT), 2012 IEEE/WIC/ACM International Conferences on , vol.3, 4-7 Dec. 2012, pp.301-305.
21. Rawat, S. ,Patil, D.R, “Efficient focused crawling based on best first search”, IEEE 3rd International on Advance Computing Conference (IACC), 22-23 Feb. 2013, pp.908-911.
22. Chang, G.J.S.; Patel, G.; Relihan, L.; Wang, J.T.-L., "A graphical environment for change detection in structured documents," The Twenty-First Annual International Computer Software and Applications Conference, 1997. COMPSAC '97. Proceedings, 11-15 Aug 1997, pp. 536-541.
23. Seung-Jin Lim; Yiu-Kai Ng, "An automated change-detection algorithm for HTML documents based on semantic hierarchies," In Proceedings. 17th International Conference Data Engineering, 2001, pp.303-312.
24. Yadav, D.; Sharma, A.K.; Gupta, J. P., "Change Detection in Web Pages," 10th International Conference on Information Technology, (ICIT 2007) , 17-20 Dec. 2007, pp.265-270.
25. Li Yue; Dong Shou-bin; Zheng Xiang; Ma Bin-Hua, "Improving navigation page detection by using DOM-based block text identification," 10th International Conference on ICT and Knowledge Engineering (ICT & Knowledge Engineering), 21-23 Nov. 2012, pp.129-134.
26. Law, Marc Teva, Nicolas Thome, StéphaneGançarski, and Matthieu Cord. "Structural and visual comparisons for web page archiving." In Proceedings of the 2012 ACM symposium on Document engineering, ACM, 2012, pp. 117-120.
27. Kuppusamy, K. S., and G. Aghila. "CaSePer: An Efficient Model for Personalized Web Page Change Detection Based on Segmentation." Journal of King Saud University-Computer and Information Sciences, 2013.
28. Varshney, N.K.; Sharma, D.K., "A novel architecture and algorithm for web page change detection," In IEEE 3rd International Advance Computing Conference (IACC), , 22-23 Feb. 2013, pp.782-787.

* * *

Student, vishnugoel90@gmail.com
Assistant Professor, vinay.arora@thapar.edu.
Thapar University,