ISSN:0975-9646

Machine Learning for Web Page Adpatation

Neetu Narwal
Asst. Prof.
Maharaja Surajmal Institute
Research Scholar, Banasthali Vidyapith

Dr. Sanjay Kumar Sharma
Associate Prof.
Banasthali Vidyapith
Rajasthan

***Abstract*—** Recent years have witnessed a drastic technological advancement in the heterogeneous display devices and they have come within the reach of individuals. But most of the websites available on Internet do not utilize the spaces available in these large screen devices as well as the small screen devices probe difficulties in adjusting the web contents on the available space. In this study, we propose a system to provide accurate and faster user perceived adaptation of the web page content. The research is in parallel to the existing W3C (World Wide Web consortium) standards and state-of-art existing web page adaptive systems, by taking into consideration web content analysis in terms of semantic coherence and block importance.

Keywords—*Web Information Retrival, Page Segmentation, Visual Blocks, web page adaptation.*

I. INTRODUCTION

World Wide Web is a collection of varied documents constituting a huge repository of information available; the massive nature of the data imposes difficulty in extracting knowledge. Analyzing such large amount of data is a challenge and provides a fertile ground for further research.

Web mining aims to discover useful information or knowledge from the web hyperlink structure, page content and usage data. Web mining uses many data mining techniques, the heterogeneity and semi-structured or unstructured nature of web data imposes difficulty in using data mining techniques directly on the web data [1]. Web mining can be broadly divided into three distinct categories, according to the type of data to be mined [Figure 1].

Web Content Mining (WCM): Web content mining is the process of extracting or mining useful information from the web page content [2]. The information available on the web sites is represented in the form of text, images, audio, video, or structured records such as lists and tables etc. Web content mining either mines the web content and provides useful information in the form of knowledge discovery i.e., cluster or classify web page according to their topic. Differentiate the main and noise content of the web page, discover useful patterns in the web page or it provides information to other tools like search engine. Further search result mining uses two different approach *agent-based* and *database*.

The *agent-based* approach to search result mining involves the development of Artificial Intelligence (AI) system that can act automatically or semi automatically to discover and organize web-based information on behalf of a particular user to provide intelligent search agents, information filtering /categorization, and personalized web agents.

The database approach focuses on database techniques for organizing semi-structured data on the web into more structured collection of resources, it uses database query mechanism and data mining techniques to analyze the information.

Web Structure Mining (WSM): Web structure mining is the process of discovering structure information from the web [3]. The structure of a typical web page can be represented as web graph, which depicts the web pages as nodes and hyperlinks as edges connecting related web pages. It makes use of different types of information to extract the structure of the web page i.e. hyperlink, document structure.

A hyperlink is a structural unit that connects a location in a web page to different location, either within the same web page or on a different web page. A hyperlink that connects to a different part of the same page is called an intra-document hyperlink, and a hyperlink that connects two different web pages is called an inter-document hyperlink [4].

Document Structure: The web page contents can also be organized in a tree like structure, using various HTML and XML tags within the page. This tree like structure is referred as Document Object Model (DOM) Tree.

WSM reveals more information than just the information contained in the web documents. For example, links from popular websites to the referred web page indicate the popularity of the web page, while links coming out of a web page indicate the richness. Hyperlink-Induced Topic Search (HITS) has been the first algorithm developed by Jon Kleinberg (1996). It is a link analysis algorithm that rates web pages and it has been widely used by search engines to indicate the richness of any web page.

Web Usage Mining (WUM): The web usage mining is used to discover interesting usage patterns from web. These patterns can further be utilized for understanding the user behavior to improve the web site accordingly. Web usage mining accepts the data from web server access logs, proxy server logs, browser logs, user profiles, registration files, user sessions or transactions, user queries, bookmark folders, mouse-clicks, scrolls, and any other data generated by the interaction of users with the web. Web usage mining uses the secondary data generated by the user's interaction, whereas web content mining and web structure mining utilize the real or primary data from the web.WUM make use of user profiles, user access patterns, and mining navigation paths for analysis. E-commerce companies make use of web usage mining for tracking customer behavior on their web sites.

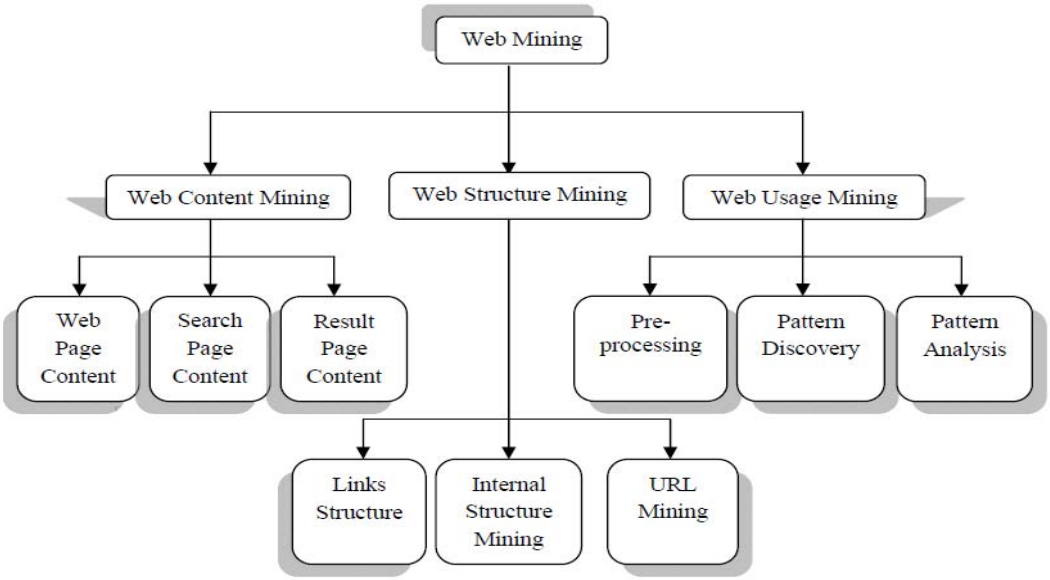

Fig 1: Classification of web mining

Web mining is similar to traditional data mining, but in data mining the data is often collected and stored in a data warehouse. For web mining, data collection is an important task, especially for web structure and web content mining, which involves scanning large number of web pages. Once the data is collected, it goes through the steps of pre-processing, data mining, and post processing as in traditional data mining [5] [Figure 2].

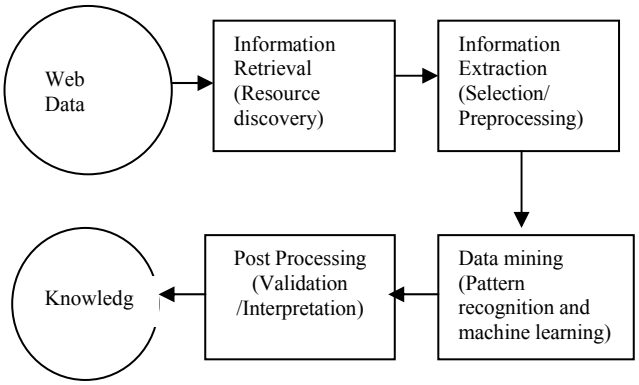

Fig 2: Steps of Web mining

The Web Mining is a field, which shows convergence to many other fields like Data Mining, Artificial Intelligence, Swarm Intelligence, Statistics etc. Data mining techniques like clustering and classification can be implemented on the data extracted from web mining to categories the web page, even swarm Intelligence can be applied to logically related contents on the web page dynamically, artificial intelligence can be used to arrange the web page content dependent on the device screen size i.e., mobile phone, palmtop, large screen display.

II. RELATED WORK

The research work aimed at web content mining focusing on web information extraction and web page adaptation. We have reviewed three major domains related to our work: Web information extraction: This is the preliminary stage of web content mining to extract information from the web page. Web content analysis: It finds useful information from the web content, like semantic coherence of the web content, noise discrimination; marking informative content based on user interest, link information etc. Web page adaptation: We studied methodologies and techniques adopted by the researchers for adapting the web content on the large and small scale device, and for user perceived filtering and final adaptation of the web content.

Web Information Extraction

Researchers have been suggested database techniques for building specialized program called wrapper for web information extraction, which maps the data of interest into specific format. The wrapper programs are categorized on the basis of human interaction with the system as manual, semi-automatic and automatic. Some of the commonly used wrapper tools are *rapier, stalker, whisk, softmealy*. In-order to extract the semantic content of the web page; various page segmentation algorithms have been developed in the past.

K.S. Kuppusamy et. al. [6] proposed a model to block the web content at a fine-grained level they suggested that instead of completely blocking the page it would be efficient to block only those segments which holds some informative content. V.R. Palekar et. al. [7] presented an approach that utilizes the visual features of the web pages to implement deep web data extraction, including data record extraction and data item extraction. Bernhard Krüpl-Sypien et. al. [8] proposed a model based on gestalt theory principles, it is implemented with semantic web techniques by considering the visual appearance of a web page. Page is viewed as a collection of gestalt figures each representing a specific design pattern. Jinbeom Kang et. al. [9] suggested

repetition-based page segmentation (REPS) algorithm which uses the repetitive tag patterns called key patterns from the DOM tree structure of the web page and generate virtual nodes to segment the web page.

Chia-Hui Chang et. al. [10] surveyed major web data extraction techniques and compared the techniques using three dimensions: the task domain, the automation degree, and the techniques used. Jie Zou et. al. [11] proposed html web page segmentation algorithm and applied it on the online medical journal, articles (regular HTML and PDF-Converted-HTML files). For a given article, a zone tree is generated by combining DOM tree analysis and recursive X-Y cut algorithm with other visual cues, such as background color, font size, and font color etc., zone tree is then segmented into homogeneous regions.

Xin Yang et. al. [12] suggested the partitioning of the web page into rectangular segments called blocks, by utilizing the visual and layout information of the web page. Deng Cai et. al. [13] suggested the vision based page segmentation (VIPS) algorithm for partitioning the web page into blocks. Their study was based on the analysis of the web page and page block relationship using link structure and page layout analysis and they constructed a semantic graph where each node represents a visual block of the web page. Xiao Dong Gu et. al. [14] proposed an automatic top-down tag-tree independent approach to detect the web content structure by simulating the web page layout based on vision.

Web content is analyzed by using machine learning and rule based algorithm to provide useful information regarding web data that can be utilized for various applications. Some of the domains are:

1. News Filtering.
2. Web Content Adaptation.
3. Spam Detection.
4. Noise Removal.
5. Search Result Mining.
6. Topic Specific Link Analysis.
7. Focused Crawling.
8. Analysis of Web Topical Structure.
9. Rank Analysis.
10. Contextual Advertising.

Web page comprises of informative content as well as non-informative content called noise like advertisements, links to different web page, navigational information etc. Researchers have worked in the area of noise detection and elimination from the web page.

N. Pappas et. al. [15] presented a unified approach to perform segmentation of the web pages and noise removal. They utilized the local features of the web page to find diversity and region accuracy to discriminate noise from the informative content. Thanda Htwe [16] proposed the mechanism to eliminate multiple noise patterns in web page to reduce irrelevant data. They applied case-based reasoning technique to detect multiple noise patterns in current web page and also present back propagation neural network algorithm for the same.

Jing Li. et. al. [17] developed web cleaner system for eliminating noise blocks from web pages for the purpose of improving the accuracy and efficiency of web content mining, and identified the importance of each block using Naive Bayes Classification. Andre L. Carvalho et. al. [18] proposed methodology that make use of web graph to identify the noisy links and shown that the removal of noisy links improves the web-page classification performance.

S. Debnath et.al. [19], designed and implemented four algorithms, content extractor, feature extractor, K-feature extractor and L-extractor that identify the primary content block of a web page. L. Yi et. al. [20] proposed noise elimination technique, based on the observation that noisy blocks usually share some common contents and presentation styles. They represented the web page as style tree, which captures the common presentation style and the actual contents of the web page, and compared web pages across a website to discriminate noise from the main content. S. S. Bhamare et. al. [21] presented a survey on web page noise cleaning for web mining; they identified the challenges and issues in this area.

Research has been done in the area of analyzing of the informative content of the web page which can be utilized for the purpose of extracting subject related information, news filtering, summarization, content adaptation etc.

Jeff J. S. Huang et. al. [22] introduced a concept of coherence set and proposed an algorithm to automatically identify and detect coherence set by comparing similarity between adjacent presentation groups. Ruihua Song et. al. [23] proposed block importance model which assign importance values to different blocks, the block features including spatial features and content features are used to train Support Vector Machine (SVM) and Neural Network methods to learn general block importance models. Shian Hua Lin et. al. [24] proposed a system called Info-Discoverer, which partitions web page into several content blocks and based on the occurrence of the features (terms) it calculates entropy value that is compared with entropy threshold for partitioning block into informative and redundant block.

Web Page Adaptation

Web Page adaptation has been the area of research since last decade, with the availability of numerous devices for web page viewing like laptops, windows workstations, iPhones, iPads, android phones with touch input, scroll wheels, keyboards, voice input, devices with pressure sensitivity, smart watches, toasters and refrigerators, and many more, various researchers have discussed the need of system for adapting the web page based on screen size [25, 26]. Ethan Marcotte has coined a new term for the web sites designed for varied range of display sizes as responsive web sites [27]. The basic idea of responsive web design is that a website should respond to the device it's being viewed on. In broad terms, responsive web site must perform the below mentioned task:

1. Adapting the layout to suit different screen sizes, from widescreen desktops to small phones.
2. Resizing images to suit the screen resolution.

3. Serving up lower-bandwidth images to mobile devices.
4. Simplifying page elements for mobile use.
5. Hiding non-essential elements on smaller screens.
6. Providing larger, finger-friendly links and buttons for mobile users.

Mobile devices differ in network bandwidth, processing power, storage, energy restrictions and format handling capabilities compared to desktop PCs. These restrictions mean that there must be some way to adapt the current web content to heterogeneous mobile devices and a method to author web content in a device independent way [28].

The process of web content adaptation can be performed on any of the three locations where each has its own advantage and limitation:

1. Client Side Adaptation
2. Server Side Adaptation
3. Intermediary HTTP proxy server that exists solely for the purpose of providing these transformation services.

Michael Nebeling et. al. [29] presented a study depicting the approximate percentage of screen space being utilized for main content and noise content of the web page in the web sites. May H. Riadh et. al. [30] presented a web content adaptation system for mobile devices. The system enables the presentation of web content by considering the problem of small screen display of mobile computing devices, also independent-device access to web content is considered.

Nobuo Funabiki et. al. [31] proposed a web-page layout optimization method for multi-modal browsing sizes. It dynamically changes box locations and font sizes by switching CSS files for different browsing sizes, so that the main content can be accessed without the screen scroll operation even at a small size whereas the blank space is avoided at a large size. Rick C.S. Chen [32] suggested fuzzy reasoning to create a functionality sense based content adaptation (FSCA) mechanism protecting semantic coherence at the time of adaptation; they introduce relevance of functionality (ROF) to quantitatively represent the similarity intensity between two presentation objects (groups).

Jaing He et. al. [33] proposed Xadaptor, which provides an extensible systematic and adaptive content adaptation approach that can be applied to diverse content types and wide variations of devices. It adopts a rule-based approach to achieve flexibility and extensibility in content adaptation. Stephen J.H. Yang [34] used JESS (Java expert system shell) to design and implement dynamic adaptation strategies to direct the transformation process. Adaptation strategies are designed to improve rule base efficiency by dynamically linking the rule based on user perceived context change.

Zhigang Hua et. al. [35] suggested an adaptive system called MobiDNA for serving dynamic content in mobile computing environment, the remote web server generates the dynamic content and transmits it over wireless network and then it is adapted for display on small screens, they integrated the web content adaptation algorithm with caching strategy to improve the performance.

Orkut Buyukkokten et. al. [36] Introduced five methods for summarizing parts of Web pages on handheld devices, such as personal digital assistants (PDAs), or cellular phones. Each Web page is broken into text units that can each be hidden, partially displayed, made fully visible, or summarized.

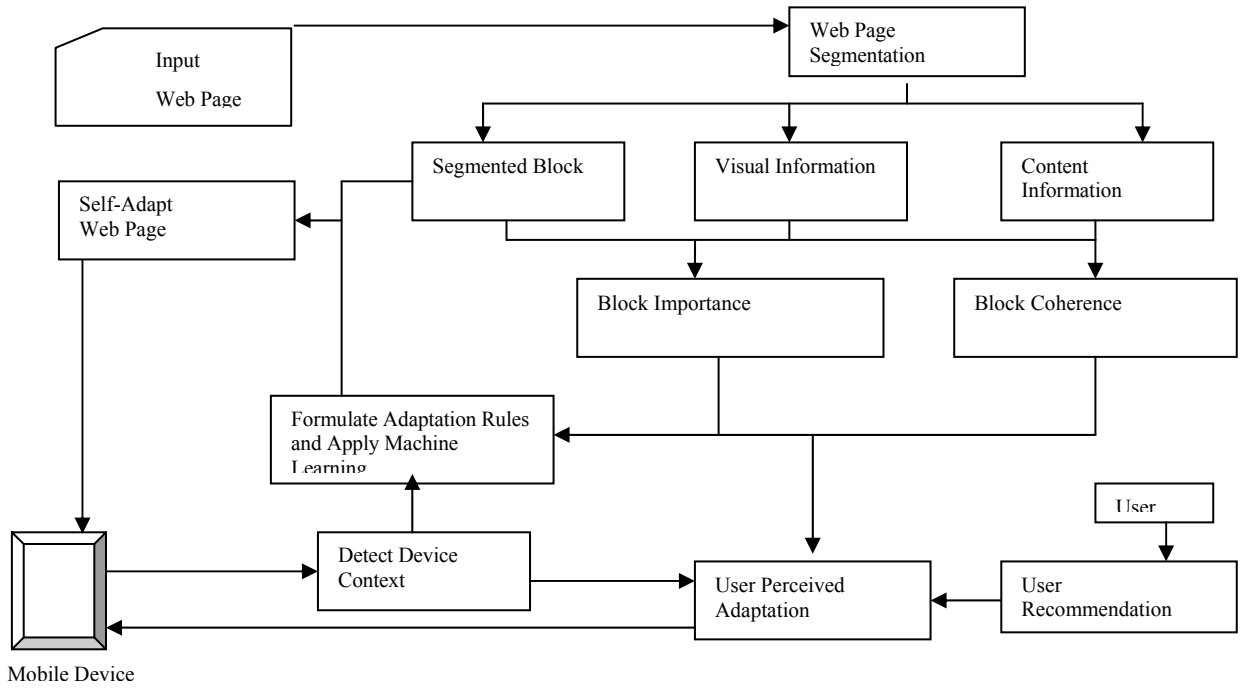

Fig : Suggested Methodology to be adopted for web page adaptation for mobile devices

III. BLOCK IDENTIFICATION MODEL

Our research is focused in the area of web content mining with special emphasis on web page adaptation and making the website responsive in accordance to the device view-port size. We propose a methodology to develop a tool that can perform user perceived web page adaptation according to the device screen size. The proposed system comprises of six steps:

Step1: Web page segmentation

The web page segmentation module partitions the web page based on semantic and visual characteristic to provide various visual blocks Gi, where the value of i range from 1 to n, where n is the total number of blocks extracted from the web page S. Union of all these blocks combined together must provide the complete set of information from the web page.

The Document Object Model (DOM) API is a set of library functions that is used to access the web page as tree structure. It can be used to retrieve, manipulate and delete the web content dynamically. By making use of DOM API web page is traversed in top-down or bottom-up manner to perform segmentation, the nodes are further merged or segmented to derive the valid visual blocks of the web page.

The final output is a set of valid nodes covering all the visual blocks of the web page with no redundancy. The web page segmentation module after extracting the visual blocks of the web page, provide three outputs:

1. The visual blocks extracted from the web page,
2. The visual information related to the web page like x-y location, height, width, font-size, color, background color etc.
3. The content related information like block content type (image, text, hyperlink etc.), block text length, count of internal links and count of external links.

Step2: Calculate block importance

Block Importance is a measure that describes the relevance of each block in context to the web page. Once the previous module has extracted the visual blocks, the machine-learning (ML) algorithm is applied to obtain the block importance. Each ML technique requires specific feature set, which is prepared from the data received by the web page segmentation module.

Every block has specific features that are used to build feature vector of each block. These feature sets are passed to the algorithm for appropriate rule induction and formulation. Block importance is useful in deciding the relevant and non-relevant portion of the web page and hence useful in the user perceived web page adaptation.

Step 3: Calculate block coherence

Block Coherence is a measure for finding the similarity between two adjacent blocks in the web page and entropy measure of two adjacent blocks are utilized for deciding the probability of keeping them together in the final adapted web page.

Machine learning algorithm is applied to measure the block coherence. The ML algorithm takes visual and content features obtained from the previous web page segmentation module as input for rule induction and formulation as done in the previous module for obtaining block importance.

Step 4: Detect device configuration

This module captures the configuration of the device i.e., screen size, network bandwidth, device capabilities and provide it as input to the main module where content adaptation occurs. Device configuration is the basis on which adaptation rules are formalized for user perceived web content adaptation.

Step 5: Formulate adaptation rules and apply machine-learning techniques

This module formulates adaptation rules by making use of Machine Learning (ML) techniques. ML appropriate feature set is prepared from the data received by the previous modules: block coherence, block importance, visual feature and content features. The feature set is passed as input to the algorithm to perform rule induction. The perceived rules are translated into media queries and added to the cascading style sheet (CSS) file. CSS file is then linked to the web page to provide final adaptation according to device context.

Step 6: User perceived adaptation

User perceived web page adaptation modules takes one time recommendation from the user like interest of the user in various subjects, noise filtering etc. This information is stored in the database for future use. The module use this information to calculate the similarity measure (cosine) of the recommended text and the text content of each visual block extracted from the previous module. The blocks having similarity measure above the threshold are selected and rearranged according to the device configuration to provide user perceived filtered web content. The block diagram of our proposed system is given in figure 3.

V. APPLICATION OF BLOCK IDENTIFICATION

The block identification plays a significant role in various web applications. The output of the model can be utilized for web content personalization, content segregation, search engine crawlers, viewing the web page on small screen device etc.

Block identification can be utilized for topic specific search where user is interested in finding the useful content related to any topic from different web site. The main content from different web sites can be clubbed and displayed to the user.

Another useful application of block identification is displaying selective content of web site on small screen devices. Due to limited screen space, main content and internal links information is sufficient to be displayed to the user.

VI. CONCLUSION

In this paper we proposed a system to extract visual blocks from web page, classify the blocks using machine learning techniques and to segregate informative and non-informative content efficiently. Finally the system filters the informative content and provide accurate and faster user perceive adaptation of content on the mobile devices. Further research work can extend this to incorporate evolutionary algorithms to web page layout optimization.

REFERENCES

[1] Bing Liu, Web Data Mining – Exploring hyperlinks, Contents and Usage Data, Springer 2007.

[2] O. Etzioni, The World Wide Web: Quagmire or gold mine, Communications of the ACM, 1996, pp 65-66.

[3] R. Kosla and H. Blockeel, Web Mining Research: A Survey.SIG KDD Explorations. Vol. 2, pp 1-15, 2000.

[4] J. Srivastava, P. Desikan P and V Kumar, ―Web Mining-Accomplishment & Future Direction, 2004.

[5] S. K. Pal, V. Talwar, P. Mitra, Web Mining in Soft Computing Framework: Relevance, State of the Art and Future Directions, IEEE Transactions On Neural Networks, Vol. 13, No. 5, September 2002.

[6] K.S.Kuppusamy1 and G.Aghila, A Personalized Web Page Content Filtering Model Based On Segmentation, International Journal of Information Sciences and Techniques (IJIST) 2012.

[7] Palekar V.R., Ali M. S. And Meghe R., Deep Web Data Extraction Using Web-Programming-Language-Independent Approach, Journal of Data Mining and Knowledge Discovery, 2012 pp 69-73.

[8] Bernhard Krüpl-Sypien, Ruslan R. Fayzrakhmanovy, Wolfgang Holzinger, A Versatile Model for Web Page Representation, Information Extraction and Content Re-packaging, ACM Digital Library, 2011 pp 129-138.

[9] Jimbeom Kang, J. Yang, J. Choi, Repetition-based Web Page Segmentation by Detecting Tag Patterns for Small-Screen Devices, Consumer Electronics, IEEE Transactions on Vol 56, Issue 2, May 2010, pp 980-986.

[10] Chia Hui Chang, M. Kayed, M. R. Girgis, K. Shaalan, A Survey of Web Information Extraction Systems, IEEE Transactions on Knowledge and Data Engineering VOL. 18, NO. 10, OCTOBER 2006.

[11] Jie Zou, Daniel Le and George R. Thoma, Combining DOM Tree and Geometric Layout Analysis for Online Medical Journal Article Segmentation, Proceedings of 6th ACM/IEEE –CS Joint Conference, JCDL 2006.

[12] Xin Yang, P. Xiang, Y. Shi, Semantic HTML Page Segmentation using Type Analysis, International Symposium of Computing and Application, IEEE Explore 2006.

[13] Deng Cai, S. Yu, and J. R. Wen., VIPS : a Vision-based Page Segmentation Algorithm. Microsoft Technical Report (MSR-TR-2003-79), 2003.

[14] Xiao- Dong Gu, Jinlin Chen, Wei-Ying Ma and Guo-Liang Chen, Visual Based Content Understanding towards Web Adaptation, Springer 2002.

[15] N. Pappas, G. Katsimpras, E. Stamatos, Extracting Informative Textual Parts from Web Pages Containing User Generated Content, ACM 2012.

[16] Thanda Htwe, Cleaning Various Noise Patterns in Web Pages for Web Data Extraction, International Journal of Network and Mobile Technologies, 2010, pp 74-80.

[17] Jing Li and C.I. Ezeife, Cleaning Web Pages for Effective Web Content Mining, Database and Expert System Applications, Springer Lecture Notes in Computer Science, Volume 4080, 2006, pp 560-571.

[18] Andre L. Carvalho, A. P. Chirita, E. S. Moura, P. Calado, W. Nejdl, Site Level Noise Removal for Search Engines, Proceedings of 15th International Conference on World Wide Web, WWW'06 ACM, pp 73-82.

[19] S. Debnath, P. Mitra, N.Pal and C.L. Giles, Automatic Identification of informative sections of web pages, IEEE Transaction on Knowledge and Data Engineering, 2005, pp 1233-1246.

[20] L. Yi, B. Liu and X. Li, Eliminating noisy information in Web pages for data mining. ACM KDD-2003, pp 296.

[21] S.S. Bharmare, Dr. B.V. Pawar, Survey on Web Page Noise Cleaning for Web Mining, International Journal of Computer Science and Information Technologies (IJCSIT), 2003 pp 766-770.

[22] Jeff J.S. Huang, Stephen J.H. Yang, Zac S.C. Chen, Frank C.C. Wu, Web Content Adaptation for mobile devices: A fuzzy based approach, A Knowledge Management and E Learning: An International Journal, Vol. 4, 2013.

[23] Ruihua Song, Haifeng Liu, Ji-Rong Wen, Wei Ying Ma, Learning Important Models for Web Page Blocks based on Layout and Content Analysis, SIGKDD 04, 6(2): pp 14-23.

[24] Shian Hua Lin, Jan Ming Ho, Discovering Informative content blocks from web documents, SIGKDD ACM, 2002, pp 588-593.

[25] M. Czerwinski, G. Smith, T. Regan, B. Meyers, G. Robertson and G. Starkweather, Toward Characterizing the Productivity Benefits of Very Large Displays, Microsoft Research, One Microsoft Way, Redmond, WA, USA, in Interact, 2003, pp. 9-16.

[26] J.H. Goldberg, and J.I. Helfman, Evaluating User Expectations for Widescreen Content Layout, Paper presented at the Usability Professionals' Association Conference, Austin, TX, USA. (2007).

[27] Ethan Marcotte, Responsive Web Design - A Book Apart, New York, 2011.

[28] J. P. LIYANAGE, Automatic Reauthoring of Web Pages for Small Screen Mobile Devices, University of Colombo School of Computing, 2002.

[29] Michael Nebeling, F. Matulic, M. C. Norrie, Metrics for the Evaluation of News Site Content Layout in Large-Screen Contexts, Proceedings of SIGCHI Conference on Human Factors in Computing Systems, ACM, 2011, pp 1511-1520.

[30] May H. Riadh, Akram M. Othman, International Journal of Computer Applications (IJCA), 2011.

[31] Nobuo Funabiki, J. Shimizu, M. Isogai, T. Nakanishi, An Extension of the Web-page Layout Optimization Method for Multimodal Browsing Sizes, Network Based Information Systems (NbiS) 13th International Conference, 2010, pp 139-146.

[32] Rick C.S. Chen, Stephen J.H. Yang, Jia Zhang, Enhancing the precision of content analysis in content adaptation using entropy based fuzzy reasoning, Elsevier: Expert Systems with Applications, 2010.

[33] Jaing He, Tong Gao, Wei Hao, I-Ling Yen, Farokh Bastani, A Flexible Content Adaptation System Using a Rule Based Approach, IEEE Transaction, 2007.

[34] Stephen J.H. Yang, Norman W.Y. Shao, Enhancing pervasive web accessibility with rule based adaptation strategy, Elsevier, Expert Systems with Applications, 2007.

[35] Zhigang Hua, Xing Xie, Hao Liu, Hanqing, Wei-Ying Ma, Design and Performance Studies of An Adaptive Scheme for Serving Dynamic Web Content in Mobile Computing Environment, IEEE Transactions on Mobile computing, Dec 2006.

[36] Orkut Buyukkokten Hector Garcia-Molina Andreas Paepcke, Seeing the Whole in Parts: Text Summarization for Web Browsing on Handheld Devices, ACM 2001, pp 652-662.