Chapter 2

Literature Survey

In this chapter, reviews of the published research work related to this work have been presented. A detailed review of web mining and their techniques along with the experimental observations which are related to the objective of this thesis have been reviewed and discussed. Lastly, the research works that are closely related to the thesis work are discussed in detail and the research gaps are identified from their contributions.

2.1 Prior Researches on Web Content Mining

World Wide Web contains the huge reservoir of information and provides the fertile ground for research work. Web mining research is a collaboration of various fields combining databases, information retrieval, machine learning technique, natural language processing etc.

**Definition:** Web mining aims to discover useful information and knowledge from web hyperlink, page content, and usage data. Web mining uses many conventional data mining techniques, but it is not purely an application of traditional data mining owing to the semi-structured and unstructured nature of the web data [51].

Lots of research have been done in the field of web mining and developed many algorithms and techniques. This area of research is prominent from last few decades and still growing due to tremendous amount of information available on the web and involvement of different work groups with Internet like e-commerce, banking, stock, education, social networking etc. Web mining is categorized into three areas based on which part of the web is to be mined [88]: Web structure mining, Web usage mining and Web content mining.

Web structure mining attempts to realize the underlying hyperlink structure of the web. It is used to categorize the web page and to generate information related to relationship and similarity between different web sites [11].
Web usage mining extracts valuable information from the secondary data derived from the user interactions with the web i.e. sessions, behavior pattern [14].

Web content mining aims at discovering knowledge from the information available on the web page. The web content encompasses variety of data i.e., text, images, audio, video, flash etc. which makes the data on Internet highly unstructured.

Web content mining in some measure related to data mining and text mining. As various data mining techniques can be applied for web content mining and the web content generally comprises of text. However, it is also relatively different from data mining because web data are primarily semi-structured and/or unstructured, while data mining deals predominantly with structured data and similarly the text mining focuses on unstructured texts. Web content mining thus requires novel approach for utilization of data mining and/or text mining techniques.

Due to the heterogeneous and un-structured nature of web data, automated discovery of knowledge still presents a challenge. Some of the problem areas where researchers have contributed towards web content mining are specified here.

1. **Data/Information extraction**: Its purpose it to extract structured data from web page, such as products comparison and search results. Many researchers have developed wrapper software that allows extraction of data using automatic, semi-automatic and manual methods [19, 34, 47].

2. **Web information integration**: Although the web contains a huge amount of data, few websites represents similar information in different forms. In this context there is a need to identify semantically similar data content and has many practical applications.

3. **Opinion extraction**: Opinion mining is a very challenging area which gathers data from online opinion sources like customers review of the products, discussion forums, online blogs and chat rooms. It provides opinion mining from the data and it is of great substance for planning marketing strategy, product benchmarking etc. Many researchers have worked in this area and provided some valuable techniques and applications.

4. **Knowledge synthesis**: Concept hierarchies or ontology of a web site has been used in many applications. However, it is difficult and time consuming to generate them manually. Researchers have worked in the area of automatic
knowledge synthesis from the web data. Its main application is to integrate and comprehend the information on the web to provide a coherent view of the topic.

5. *Segmenting web pages and detecting noise:* Partitioning of the web page into blocks is required by various web applications that need only the relevant content of the web page. The content like advertisements, navigation links, copyright notices are considered as noise. A number of interesting techniques have been proposed lately for web page segmentation and noise detection [8, 10, 44].

All these tasks put forward a major research challenges and their solutions have immediate real-life applications.

This dissertation draws attention on three main aspects of web content mining – web information extraction, web content analysis and web page adaptation. The related work in the field of web content extraction and other intelligent systems for web content classification, adaptation and web personalization are discussed further.

1. *Web information extraction:* It is the primary step in web content mining, which extracts information from the web page.

2. *Web content analysis:* It analyze and extract 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.

3. *Web page adaptation:* It uses different methodologies and techniques for adapting the web content on large or small scale devices.

4. *Soft Computing Techniques for web content mining:* The various soft computing and hybrid techniques that are generally used by the researchers for web content mining.

5. *Web Personalization:* Web Personalization is an important aspect of Web 2.0, where the web page can be personalized based of the characteristics, actions, intentions and any other parameter associated with the user.
2.1.1 Web Information Extraction

Earlier researchers have suggested some of the database techniques for building a specialized program called wrapper for web information extraction. A wrapper is a specially designed procedure for extracting content from information source and delivers the content of interest in a descriptive representation. A wrapper accepts a query from the user applies a set of extraction rules and returns the relevant information [20]. To extract information from several independent sources, libraries of wrappers are needed. Web is ever changing hence wrappers need to be adjusted dynamically.

The wrappers are generated using manual, semi-automatic and automatic approaches. The manual generation of a wrapper often involves the writing of static code. The developer first understands the structure of the web document and then translates it into program code. The manager of multiple information sources system is the first manual wrapper constructed at Stanford IBM. The goal of the system is to provide methods for accessing data in an integrated fashion from multiple sources, and ensure the information consistency.

Semi-automatic approach makes use of support tools to design the wrapper. Some of these approaches take user input regarding the portion of the web document that need to be extracted. Automatic approach utilizes the power of machine-learning techniques to build wrappers that extract information from the web by analyzing the content without human intervention. A wrapper in this category ranges from simple to relatively complex.

ShopBot [19] is a shopping wrapper, designed to extract information from different web vendors and represent a comparative analysis of a product. The algorithm utilizes the tabular format of the web page content available in the vendor web sites. Information is extracted from different vendor’s website by making use of heuristic search, pattern matching and inductive learning techniques.

The Wrapper Induction Environment [47] is a tool for developing wrapper software. It works on structured data in tabular format in the HTML documents. Soft mealy [34] is wrapper software that extracts data from semi-structured web pages using non-deterministic finite automata and an inductive generalization algorithm. It is used to discover the contextual rules from training samples. Gogar et. al. [26] presented self
learning wrapper software. The wrapper uses Neural Network to extract information from a previously unseen web page and does not need any site specific initialization. They proposed a method for spatial text encoding that allows encoding of visual and textual content of a web page into a single Neural Network.

Wrapper is a specific approach to extract information from the web, whereas another class of technique namely web page segmentation is a generalized approach of information extraction from the web page. The web page segmentation aims at partitioning a web page into blocks that are cohesive and depict the presentation logic used by the web page designer.

There are different approaches used by researchers for web page segmentation. Some of the most popular approaches are:

1. Fixed length web page segmentation
2. DOM based page segmentation
3. Vision based page segmentation
4. Combined/Hybrid method

The *heuristics-based* web page segmentation is the most popular technique where researchers have contributed different heuristic strategies and formulated rules for segmentation. Cai D. et al. [8] used heuristic rules to extract the structural and visual properties of each block of a web page. Their procedure recursively partitions the larger blocks into smaller ones using a top-down approach. However, it is difficult to build generic rules that can be applied on different type of websites. Kohlschutter C. et al. [44] employed text density as a clue to perform segmentation of a web page.

Chakrabarti D. et al. [10] proposed a graph-theoretic approach for web page segmentation. Considering the segmentation as a minimum cut problem on a weighted graph where the nodes are the DOM tree nodes and the edge weights is the cost of placing the end nodes in the same segment or different segments. A learning based method is used to determine the weights of edges. Different from Cai D. et. al. other researchers Kohlschutter C. et. al. and Chakrabarti D. et. al. obtain a flat segmentation of a web page without knowing the hierarchical structures of the web page.

Kuppusamy K.S. et al. [46] proposed a model to segment the web page at a fine-grained level and suggested that one should consider only those blocks which contain
some informative content. Palekar V.R. et. al. [64], presented an approach that utilizes the visual features of the web pages to perform deep web data extraction.

Kang J. et. al. [41], 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.

Zou J. et. al. [88], proposed html web page segmentation algorithm and applied it on the online medical journal and medical articles. In their methodology they prepared Zone tree by using recursive X-Y cut algorithm and DOM tree analysis along with some other visual features such as background color, font size, and font color etc. The zone tree is then segmented into homogeneous regions. Yang X. et. al. [83] suggested the partitioning of the web page into rectangular segments called blocks, by utilizing the visual and layout information of the web page.

Cai D. et. al. [8] suggested the vision based page segmentation (VIPS) algorithm for partitioning the web page into blocks. Their study is 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. Gu X.D. et.al. [29] proposed an automatic top-down tag-tree independent approach to detect the web content structure by simulating the web page layout based on vision. Swezey et. al. [76] proposes a web page segmentation algorithm based on title blocks. The web pages is divided into minimum number of blocks and are classified based on the features of the block as title or non-title block. The title blocks are used as separator and non-title blocks are assembled into web content blocks.

Safi et. al. [67] proposed a framework to enhance the web accessibility for visually impaired people by providing them a first glance web page overview. They suggested a hybrid segmentation algorithm to provide easy navigation of web page. They transformed the layout of the web page into a coarse grain structure, which is transformed into vibrating pages using graphical vibro-tactile language.

### 2.1.2 Web Content Analysis

Identification and analysis of the web content is used for different web applications. One of the applications is to re-organize a web page so that it can be properly displayed on devices with small screen size [4, 31, 48]. Other important application being the main content identification and filtering which is a very useful for visually
impaired people who can focus only on the relevant content of the web page by making use of their screen readers [49]. The applications like link analysis, web document indexing, pseudo-relevance feedback, duplicate content detection, web page classification and content change detection can be made more effective with the accurate classification of web content [55, 73]. Other common application of web content analysis is listed as under:

1. News Filtering.
2. Web Content Adaptation.
5. Search Result Mining.
6. Topic Specific Link Analysis.
7. Focused Crawling.
9. Rank Analysis.

Web page normally comprises of informative content as well as non-informative content also called as noise i.e., advertisements, links to different web page, navigational information, copyright etc. Researchers have worked in the area of noise detection and exclusion from the web page.

Pappas N. et.al. [65] 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. Htwe T. [35] proposed the mechanism to eliminate noise patterns from a web page by eliminating irrelevant and redundant data. They applied case-based reasoning technique to detect multiple noise patterns in web page and also used back propagation neural network algorithm for the same.

Li J. et. al. [51] developed web cleaner system for eliminating noise blocks from web pages to improve the accuracy and efficiency of web content mining, and identified the importance of each block using Naive Bayes Classification. Carvalho A. L. et. al.
proposed methodology that makes use of web graph to identify the noisy links and showed that the removal of noisy links improves the web-page classification performance.

Debnath S. et.al. [16] designed and implemented four algorithms i.e., content extractor, feature extractor, K-feature extractor and L-extractor that identify the primary content block of a web page. Yi L. et. al. [84] proposed noise elimination technique based on the theory that the noisy blocks usually share some common structure and presentation styles. They represented web page style tree, which captures the common presentation style, actual contents of the web page, and compared web pages across website to discriminate noise from the main content.

Huang J. S. et. al. [37] introduced a concept of coherence set and proposed an algorithm to automatically identify and detect coherence set by comparing similarity between adjacent presentation groups. Lin S. H. et. al. [54], 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. Bhamare S.S. et. al. [5] presented a survey on web page noise cleaning for web mining; they identified the challenges and issues in this area. Nebeling M. et. al. [57] 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.

2.1.3 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, personal computers, mobile phone, tablets, palmtops, smart watches, and many more. Various researchers have discussed the need of system for adapting the web page based on screen size [15, 27]. Marcotte E. has coined a new term for the web sites designed for varied range of display sizes as Responsive web sites [56]. The basic idea of responsive website design is that it should respond to the device being viewed on. In broad terms, responsive web site must perform the below mentioned task:

1. Adapting the layout according to different screen sizes, ranging from widescreen desktops or personal computers to small screen mobile devices.
2. Resizing images for better adaptation according to device resolution.

3. Serving up lower-bandwidth images to mobile devices.

4. Simplifying page elements or filtering content so that they can be viewed conveniently on mobile browser.

5. Hiding irrelevant content for smaller screen devices.


Mobile devices differ in network bandwidth, processing power, storage, energy restrictions and format handling capabilities compared to desktop and personal computer. 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 [53].

The web content adaptation process can be performed on any of the three locations of the system, 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.

May Riadh M. H. et. al. [66] presented a web content adaptation system for mobile devices. The system adapts the web content based on the small screen of mobile devices and also provides device-independent access. Ahmadi et. al. [1] presents a novel approach that automatically adapts a web page for mobile presentation. Their approach comprises of two steps: first it finds boundaries among different information blocks and then generate an adaptive layout for mobile devices. The web page segmentation produces a block tree, which is used for generating adaptive layout, which places closely related information in proximity and minimize scrolling. Funabiki N. et. al. [22] proposed a web-page layout optimization method for multimodal Internet browsing devices. The web page is designed using box layout where the system dynamically changes box location and other formatting features by switching CSS files designed for different devices, 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.
Chen R. C. S. [12] used fuzzy reasoning to perform a functionality sense based content adaptation (FSCA) thereby protecting the semantic coherence of objects at the time of content adaptation; they introduce relevance of functionality (ROF) to quantitatively represent the similarity between objects. Jaing H. et. al. [39] designed a tool named Xadaptor that provides a systematic and adaptive content adaptation technique. Tool applies a set of content adaptation rules on varied content types and for multi modal devices.

Yang S. J. H. [82] 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. Hua Z. et. al. [36] 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 its performance. Wei et. al. [78] presents the architecture, design and implementation of a tool named website mobilizer. It semi-automatically converts existing web pages to web pages suitable for mobile viewing. The mobilizer extracts the component hierarchy of the page using web crawler techniques, analyzes the structure of the web page, and then converts it to the mobile version using a combination of techniques that includes link analysis, ranking algorithms based on component content, and interactive removal of irrelevant content.

Augustine et. al.[3] proposed a new method for web content adaptation for mobile devices and provides a condensed view of an adapted web page. They used depth-first algorithm for categorization of HTML objects and then select the sequence of blocks to be displayed on a mobile device. They evaluated the usability of their adapted contents against other deployed systems. Kao et. al.[40] proposed a platform for running web application on mobile devices. With the help of their platform the application or web developer can develop and publish offline services. Like web content adaptation, synchronization, offline services etc. Buyukkokten O. et.al. [7] suggested five different methods for summarizing the content of web pages on handheld devices, such as mobile phones, palm tops, digital assistant. Each web page
is broken into text units that can be displayed in one of the given modes i.e., hidden, partially displayed, made fully visible, or summarized.

### 2.1.4 Soft Computing Techniques for Web Content Mining

Machine learning techniques are widely used for classification and categorization. Dehankar et.al.[17] in their work presented a web page classification system that make use of Aproiri Algorithm and Naïve Bayes Classifier. The keywords entered by user in search engine are picked by Aproiri algorithm to get the links between associated word and the page content. Then Naïve Bayes Classifier is used to calculate the probability of each word with respect to the class. Kim et. al.[43] proposed a web page classification system that make use of meta-tag available in the web page for classification. Each web page comprises of meta-tag that stores the terms related to the web page, hence used for accurate classification of web page. Shukla et. al. [69] proposed an automatic genre based web page classification system. The system based on the content of the web page determines whether it contains an advertisement or not. They used key features of the content and based on that they define the advertisement category which is useful for numerous web applications.

Fuzzy logic is a useful tool for modeling complex real life systems, deriving useful fuzzy rules and their relationship [60] (Okamura et. al. 1998).

The hybrid of fuzzy and evolutionary algorithm have found greatest acceptance in solving variety of scientific and real life problem. Evolutionary algorithms are used efficiently with fuzzy control system to perform variety of task i.e., generation and optimization of fuzzy rule base, generation of membership function, tuning of membership function etc. Various researchers have worked with these hybrid systems.

Shill P.C. et. al. [71] designed an automatic fuzzy rule reduction system for fuzzy logic controllers (FLCs) using genetic algorithms (GAs). The schema adopted by the researcher comprises of two phases: the first phase optimizes the FLCs membership functions and second phase is the rule learning and reducing phase, this phase automatically formulate the fuzzy rules as well as determines least set of fuzzy rules that are required for building the fuzzy models. In the reduction phase, the system removes the redundant rules by setting all consequent weight factors to zero and merging the conflicting rules.
Wei Juan et. al. [80] proposed a hybrid technique that utilizes the power of genetic algorithm and ant colony algorithm for adjusting fuzzy control rules. The Genetic algorithm generates initial candidate fuzzy rules which are utilized further by ant colony algorithm. Wei S. et. al. [74] used particle swarm optimization (PSO) algorithm to optimize the fuzzy control rules for the design of fuzzy controller. The result demonstrated that PSO can be applied to optimize fuzzy control rule effectively.

Ho S. Y. et al. [32] used Genetic algorithm to search for optimal fuzzy rules from the scatter feature space to design a compact fuzzy rule based system. Zhou E. et. al. [87] proposed a method for designing a fuzzy-rule-based classifier where fuzzy membership functions and the size and structure of fuzzy rules are extracted from the training data using GA. Pablo A. D. et. al. [62] also proposed an approach for automatic fuzzy rule base generation and optimization by means of Self-Adaptive Genetic Algorithm. They changed the crossover and mutation rates to avoid premature convergence and hence maintain population diversity. Hwang H. S. et. al. [38] presented a similar use of evolutionary programming for optimal generation of fuzzy rule base for modeling and control. Genetic algorithms (GAs) were applied to optimize the fuzzy membership functions or fuzzy rules separately.

Kayaa M. et. al. [42] used genetic algorithm along with clustering method to adjust the centroid of the cluster. The membership functions are generated by taking centroid as midpoint of membership functions.

Ng W. et. al. [58] also presented a genetic algorithm learning module (GALM) to optimize a fuzzy rule-based system for process control. GALM is used to design a minimum set of fuzzy if–then rules that establishes the appropriate mapping from input states to control actions of the system. With this approach, a considerable reduction in the number of rules is noticed when compared to manual design.

Ghezelayagh H. et.al. [23] also used Genetic Algorithm (GA) to optimize the fuzzy rules and fine tune membership functions used to generate neuro-fuzzy inference system for identification of the boiler system.
2.1.5 Web Content Personalization

The personalization of web refers to a collection of technologies that present the ability to reorganize, configure and manage online content rather than just viewing it. Web 2.0 describes World Wide Web sites that accentuate user-generated content, usability, and interoperability [24, 59, 74, 77]. Web 2.0 is a new version of the World Wide Web; however it focuses on how web is perceived by the user and the changes that need to be implemented in the design of web page.

A Web 2.0 web site may allow users to interact and contribute towards generation of web site content and thus creating a virtual community, where user instead of one sided viewing of the web content can redesign and reorganize the web site. Examples of Web 2.0 include social networking sites, online blogs, Wikipedia, folksonomies, video sharing sites, hosted services, web applications, and mashups.

Díaz O. et. al. [18] developed WebMakeup a browser plug-in that empowers user to re-arrange web content at wish. Users can delete irrelevant content from the web site and add new content altogether from different web site thus provide a coherent group of information for single viewing and reduced movement. This results in user driven web customization. Albanese M. et. al. [2] presented a web mining strategy for web personalization based on pattern recognition strategy that analyze and classify the content using their static and dynamic features. They conducted experiments on large number of commercial web site and demonstrated the effectiveness of their proposed system.

Shukla R. K. et. al. [70] presented a comprehensive overview of the web personalization process, techniques and highlighted their prominent features to provide support for web personalization. They discussed about the work done in this area and also about Internet companies and the implementation of web personalization systems.

Saxe R. S. [68] reviewed existing personalization techniques and analyzed the characteristics of existing database techniques applied for website personalization. They identified the ways in which active databases and data mining techniques can be combined to produce an improvement in existing website personalization and provide a scalable personalization framework.
2.2 Related Research Work

Research work done by Song R. et.al. [73] of Microsoft Asia is closely related to the work presented in this thesis, they proposed Learning Important Models for Web Page Blocks based on Layout and Content Analysis. They proposed a block importance model to assign importance values to different blocks in a web page. The Vision-based Page Segmentation (VIPS) algorithm proposed by Cai D. et. al. [8] of Microsoft Asia is used to partition a web page into blocks by analyzing the visual layout of the page and the content coherence. The spatial features and content features of each block are extracted to represent the block. They used the concept of learning by example, where the manually labeled dataset was used to train the Support Vector Machines (SVM) and neural network classifiers to create the block importance model.

In this thesis work web page segmentation algorithm is used to extract visual block and considered some new features under the category of embedded objects i.e., iframe, JavaScript element, image element, internal hyperlink and external hyperlink. We implemented our concept and find that these features have significant role in analyzing the visual block. We tested our system with well-known classifiers Neural Network, RBF Neural Network to check the feasibility of feature set in deriving the accurate classification results. We achieved the accuracy approximately 97.51% with Feed forward neural network and 99.71% with RBF neural network classifiers.

Another important research work done at Microsoft Asia by Xie X. et.al. [81] an efficient browsing of web search results on mobile devices based on block importance model. They used their previous work of learning block importance to identify the blocks and displaying them on mobile devices. In order to fit the web page into small screens, different approaches were adopted and three presentation schemes were used to display different levels of details to the mobile user. The performances of three interfaces were evaluated using user study experiments. They present a search page to the user with input box and search button, utilizing the search result of Google Search Engine, each item is shown with three different types of options

- “T” (thumbnail view with annotation)
- “O” (optimized one column view)
- “M” (main content view).
For the “T” option, the page content blocks are shown as a thumbnail to fit the screen width while they preserved the original two-dimensional layout. User can click on the thumbnail and view the enlarged view of the block individually. The thumbnail of each block is labeled with importance value that help user to identify relevant information.

Many commercial web browsers e.g. Opera, PocketIE etc. provides the functionality for re-formatting the web pages into a single column to make it fit according to available screen width. The one-column view facilitates the reader by eliminating the need of horizontal scrolling, whereas it results in comparatively large amount of vertical scrolling. In some cases user need to scroll down much more to access the main content. They proposed optimized one-column view of the webpage; where the blocks in a web page are sorted according to their importance value. Afterwards, the page is reorganized using some of the available formatting algorithms into one-column view. For the “M” view, they extracted pure text from the important blocks of a web page which are then rearranged and displayed to the user. It provides more relevant and clear results to the user.

In this thesis focus is primarily on presenting clean web page to the desktop as well as mobile user. The noise element blocks marked by classification algorithm are filtered and arranged according to their precedence in the new clean web page. The approach is similar to the “M” view of the mentioned research however this work retains the complete visual block rather than text content only. Since most of the web sites contain multimedia like images, audio, video etc as the main content and it has same relevance as text content.

The research gap identified from existing work, talks about user perceived presentation of mobile content. A new kind of web personalization is presented for the mobile user, where only the visual block that matches with user perceived query is presented to him/her. It reduces the size of web page tremendously and is easily adapted on the mobile browser with very little scrolling. It is based on assumption that the blocks which are closely related to the user query are important for the user.