Chapter 1

Introduction

The World Wide Web (WWW) since its inception in 1994 by Massachusetts Institute of Technology (MIT) and European Laboratory for Particle Physics (CERN) as an International organization laid a foundation of the Internet. The main objective of W3C’s was to establish standards for the development of the Web and provide interoperability between WWW products and specifications. Since 1995 to 2010, WWW has undergone an exponential growth in terms of amounts and variety of information, number of web users, applicability and has witnessed the collaboration of different fields on the Internet.

Today web is considered as global platform to launch new products, perform business, sale, research, marketing, education and other information sources. This resulted change in scenario how we perceive the web.

Web is a conglomeration of varied information related to diverse fields and thus forms a huge repository of information available. The massive nature of the web imposes difficulty in extracting knowledge from this reservoir. Thus web poses a challenge and provides a fertile ground for the research.

Some of the characteristics of the web that makes it more challenging are [52]:

1. **Diverse Nature**: The amount of information available on the web is huge and diverse. There exists a lot of redundant information relating to similar topic which imposes difficulty in integration.

2. **Data Format**: There exists no standards for creating websites hence data available on the web is present in varied forms e.g. structured tabular data, semi-structured web content, unstructured text and multimedia file i.e., image, audio, video.

3. **Authenticity**: Web doesn’t have any authorship hence authenticity of information is a challenging task.

4. **Noise**: The web page comprises of enlightening content along with navigational links, advertisements, copyright information, privacy policies etc. These noise data impose difficulty in analyzing the content of the web page.

5. **Scalability**: Web sites can be accessed using variety of devices running on different platforms. Scalability is also a major challenge imposed by the web.
In decades the structure and presentation style of web document has changed, it is no longer the pure information source but rather a collaboration of different non related information put forward to the web user. However the intention of web user today is to access clean information from the web site and ignore the non-relevant content. It is very easy for humans, to identify the main content from the complete web page as they can use their intellect, visual clues and layout to distinguish it from the other non related information. However, computer software is not as intelligent as human; it becomes difficult to distinguish the main content from the web page that contains noise.

Today the usage of Internet is not just restricted to personal computers, laptops etc. it has reached in the hands of users in the form of mobile devices, palmtops, tablets. Due to availability of Internet services on the mobile devices, there is global increase in the usage of mobile devices for the web access. However, these devices face a major challenge of limited screen space, less bandwidth and processing capabilities. A majority of the websites are still there to be viewed only on personal computers, desktop computer, laptops properly. Hence mobile Internet users face difficulty in browsing the web pages.

Considering these problems, there is a strong need to design technique for providing Internet user a clean web page. The thesis introduces a solution to the above question by investigating the following hypothesis:

The presented approach automatically discovers the main content blocks of the web page by using web page segmentation and classification algorithm. The main content blocks from the web page are filtered and rearranged according to device context and thereby providing web user a clean web page.

Figure 1.1 shows the home page of Delhi Haat website (www.dillihaat.in). As shown the web page is divided into blue rectangular blocks. It can be visualized that blocks numbered 3, 4, 5, 6 and 9 are more related to the title or topic of the web page. However, blocks numbered 1, 2, 7, 8 are rather unrelated content mainly advertisements, copyright, external hyperlinks etc.

It is also observed that web page comprises of not only the relevant content but also the nonrelated noise content which befits a major part of the web page structure.
Fig 1.1: a) Web page of Dilli Haat web site showing advertisements, external links, internal links with main content uploaded on 5th Nov 2014. b) Clean web page displayed on mobile phone.

1.1 Web Mining and Information Extraction

In order to extract knowledge from today’s web which is heterogeneous, unstructured and unclean is a challenge. A different set of data mining technique namely web mining is used to extract knowledge from the web data [52].

Web mining techniques are broadly categorized into three distinct categories, according to the type of web data to be mined namely [Figure 1.2]:

1. Web Content Mining
2. Web Structure Mining
3. Web Usage Mining

Web Content Mining (WCM): Web content mining is the technique of extracting useful information from the web page content [21]. The information available on the
Web sites is represented in the form of text, pictures, audio clips, video clips, or structured records such as lists and tables etc. Web content mining 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 content and noise of the web page, discover useful patterns in the web page etc. It also provides information to other tools like search engine namely search result mining. Further search result mining generally incorporate two different approaches namely agent-based and database.

The agent-based search result mining utilizes the concept of Artificial Intelligence (AI) to create a system that discover and organize web-based information automatically or semi automatically on behalf of a particular user. This approach usually provides intelligent search agents, information filtering/categorization, and personalized web agents.

The database approach on the other hand uses the concept of database management for organizing semi-structured data available on the web into more useful structured collection of data. It uses the query tools and mining techniques to analyze the data to provide useful information.

**Web Structure Mining (WSM):** Web structure mining is the technique for discovering structure information from the web page [45]. The structure of a web page can be represented in the form of graph that shows the web pages as nodes and hyperlinks as edges that connects 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 an element in the web page that direct the user to desired content which may be available on the same or different web page. A hyperlink that allows user to move to different parts within the same web page is called an internal hyperlink, and a hyperlink that allows user to connect to different web pages is called an external hyperlink [72].

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

WSM disclose much more information than just the content of the web document. For example, a hyperlink pointing in, from the popular web sites to the referred web
site signifies the popularity of the web site, whereas the hyperlinks pointing out from the web site indicates the richness or the variety of topics covered in the web page. Hyperlink-Induced Topic Search (HITS) was a popular and novice algorithm developed by Jon Kleinberg (1996) that performs link analysis and rates the web pages and has been widely used by Search Engines to indicate the popularity of any web page.

**Web Usage Mining (WUM):** Web usage mining is the technique to uncover interesting usage patterns of the web user, these patterns can further be utilized for better understanding of the user behavior and can be utilized to improve the web site accordingly. Web usage mining utilizes the data from various sources such as user access log files managed at web server, proxy server, web browser and user profile information from user registration details, analyzing user sessions etc. Web usage mining uses the secondary data generated by the user’s interaction, whereas web content mining and web structure mining uses the primary data of the web.

WUM is popularly used by various E-commerce, Social networking websites that utilizes the information of web usage mining for tracking customer behavior, understanding their patterns and trend analysis.

**Fig 1.2 Classification of Web mining**

The process of web mining is similar to traditional data mining. The first step being the identification of information source, then the information is extracted from the web source using some preprocessing and segmentation techniques. The extracted...
information is analyzed using different data mining or other relevant techniques i.e., pattern discovery, search result analysis etc.

The next step is to perform validation and interpretation of the result. Finally, the last step is the knowledge discovery [63] [Figure 1.3].

Web mining is a field which has shown convergence of various fields like Data Mining, Artificial Intelligence, Swarm Intelligence, Statistics etc. In order to integrate web intelligence into web mining, there is need to embed artificial intelligence into the web tools simultaneously eliminating the need for human intervention. A number of researchers from various fields like informational retrieval, knowledge mining, soft computing, evolutionary computing are focusing on developing intelligent systems that can effectively mine the knowledge from web.

In view of above difficulty to incorporate web intelligence into the automated web tool, various soft computing techniques have been explored for web information extraction. In this section the information extraction techniques and their applicability for web mining have been discussed.

The Information Extraction is the technique of extracting structured information from the unstructured text. Lately Information Extraction (IE) has started moving towards the web as source of textual information. Thus giving rise to new research area namely Web Information Extraction (WIE). Web Information Extraction systems differ significantly from traditional systems in terms of methodology and the outcomes. Traditional IE systems extract information from the available corpora, whereas WIE system extracts information from the huge reservoirs of web data available in the form of structured, semi-structured and un-structured information.

WIE systems face plentiful challenges as they must cope with the huge pool of information available on the web. The web data is highly unstructured and thus poses a difficulty to directly implement information extraction techniques on the web.
Unstructured nature of the web also poses other challenges like authentication, authorization, validity of data etc. Unstructured nature of web though offers certain advantage over previously used text collections i.e., *Redundancy*: Web contains redundant information from various sources, *Multiple paraphrases*: Information available on the web is likely to appear in a variety of forms, *Easy to understand language*: Same facts are likely to be repeated and restated, some of them are easier to understand than others, *Broad Coverage*: There exist numerous websites covering the same topic, hence information available on topic is wide.

### 1.2 Applications of Web Information Extraction

The Web Information Extraction (WIE) systems finds its applicability in varied application domains i.e., i) *Question Answering review* – where the answer relevant to the question is searched for the user from the web page. The web page content are tested for the probability of becoming the answer to the query. ii) *Facet search* – it returns the definition of the queried term after searching few related websites iii) *Image search* – where the relevant images are returned after analysing various images with the description provided iv) *Document Search* (web page search) – it is used by search engine application which returns the relevant web pages that matches the user request. Some other applications are *opinion mining, relevance ranking, web page content ranking, query analysis, web page crawling, web page Indexing, preparing search result, Anti-Spam* etc.

Search engine application also uses information extraction techniques to create indexed databases. They make use of specialised crawler applications that are search oriented and involves human participation which makes it a time consuming process. However human intelligence plays a significant role in avoiding the mistakes in the extraction process.

One of the vital task related to web information extraction is web page segmentation. The aim of web page segmentation is to partition a web page into blocks that are coherent in themselves thereby describing the information presentation style of the page designer. As shown in Figure 1.1, the web page is segmented into blocks marked with blue rectangles.
1.3 Soft Computing Techniques for Web Content Mining

Soft Computing techniques are an association of methodologies that works together and provides flexible information processing capability for handling real-life situations[85]. It take advantage of the acceptance for imprecision, uncertainty, approximate reasoning, and partial truth in order to achieve robust, low-cost solutions, and that is closely related to human-like decision making capability.

Some of the prominent soft computing tools like fuzzy sets, artificial neural networks (ANNs), evolutionary algorithm, genetic algorithm(GA) have been widely used by researchers for web mining. Fuzzy sets provide an appropriate framework for the task that deals with uncertainty. Neural networks (NNs) are widely used for modeling intricate functions, and provide learning and generalization capabilities. GA is a proficient search and optimization technique. These techniques are described in brief in the next section.

1.3.1 Fuzzy Inference System

Fuzzy logic has emerged from the pioneering work by Zadeh(1965) on fuzzy set theory for solving engineering control problems. The Fuzzy rule-based inference systems have become prevalent in the real life applications related to different fields [58, 60, 71]. They have proven to be highly effective in the domain with continuous, ambiguous or linguistic variables. An important component of Fuzzy Inference System are the fuzzy rules. Various techniques have been proposed for the automatic generation of fuzzy if-then rules from numerical data for classification. The research work utilizes the power of Fuzzy Inference System for web content classification.

1.3.2 Genetic Algorithm

Genetic Algorithm (GA) is a heuristic search and optimisation algorithm based on the evolutionary ideas of natural selection, survival of the fittest theory and genetics. This algorithm has been profoundly utilized to solve various optimization problems. The basic techniques of the GA are designed to simulate natural processes of genetics for evolution and they follow the principle of "survival of the fittest" laid down by Charles Darwin. The application of GA along with fuzzy classification system has been utilised for solving the problem of web content identification.
1.3.3 Artificial Neural Network

Artificial Neural Network (ANN) is one of the best machine-learning technique for solving problem that can’t be solved using conventional method. An Artificial Neural Network (ANN) is an information processing system that is capable for solving complex input-output relationships. When new input is provided to the ANN model, it produces an output similar to the closest matching training input pattern [28]. In neural network model architecture, each node at input layers receives input values, perform processing and send it to the next layer. The key feature of neural networks is that it learns the input-output relationship through training. The feedback of the neural network is utilized and the configuration is improved until the analysis of the training data reaches a satisfactory level. ANN has been used in classification of web content identification.

1.4 Objectives of the Research

The main objectives of the research work that are covered in this thesis are:

1. To study the available techniques for mobile web page adaptation.
2. To design and implement a new approach that improves the internet experience of mobile user.
3. To implement the user’s perspective of viewing the relevant website content.
4. To test the feasibility of designed technique and perform comparison with the existing research work.

1.5 Organization of Thesis

The Chapter-2 gives an extensive literature survey on web content mining and web page adaptation techniques that are directly related to the research problem. Chapter-3 describes the motivation for the research work. Chapter-4 describes the methodology used to meet the research objectives. It further describes the techniques used for web page classification and adaptation according to device context. Chapter-5 describes experiments and results of the research work and also presents a comparative analysis with the existing state of art research work. Chapter-6 gives the summary of research work, conclusions and future work.
1. Song R. et.al.[73] have used the Vision-based Page Segmentation (VIPS) algorithm proposed by Cai D. et. al. [8] of Microsoft Asia to partition a web page into blocks and obtained an XML file describing the features of each block. The algorithm used in our research works on the same line as VIPS and generates a different set of features which play significant role in this work.

2. Song R. et. al.[73] have used only the features provided by the VIPS algorithm. Their experimental result shows that Support Vector Machine (SVM) with radial basis function (RBF) kernel can achieve the best performance with Micro-F1 79% and Micro-Accuracy 85.9%. In the present work few more interesting features namely embedded objects have been utilized for classification. The result shown in Table 5.2 denotes their significance for web page block classification. The new dataset has obtained the accuracy of 99.71% for the RBF neural network classifier.

3. In this thesis the feasibility of three different classifiers are tested along with variation in the feature set. The results demonstrate that all the classifier gives accuracy more than 93.57%.

   a. In the first dataset, top 20 features out of 33 features are used as input to feed forward neural network and RBF network classifier. An accuracy of 95.51% and 99.71% is achieved respectively.

   b. In the second dataset, Shannon Information entropy value for each relative features and term entropy features are used. These four cumulative feature values are used as input to the feed forward neural network and RBF network classifier. An accuracy of 96.03% and 99.6% is achieved respectively.

   c. In the third dataset, three cumulative feature values are used to derive fuzzy rules for web content classification. The nine fuzzy rules obtained by genetic algorithm are used to build the fuzzy inference system and the FIS classifier has resulted in accuracy of 93.57%.

The classification results obtained by all three classifiers are better than the work done by Song R. et. al. [73] for web page blocks classification.

The second state of the art work by Xie X. et. al. [81] at Microsoft Asia in their paper titled “Efficient browsing of web search results on mobile devices based on block
importance model” is compared with the work done in this thesis. The following points are observed and stated as under:

1. Xie X et.al.[81] have used their previous work of visual block classification to identify the importance of each block. As discussed earlier the classification results presented in this thesis is better than their work for block classification.

2. In their work they presented three views for content adaptation for mobile phones: “T” (thumbnail view with annotation), “O” (optimized one column view) and “M” (main content view). The research work presented in this thesis works on the similar line as their third view. They have extracted only the text content of the main blocks, since the significance of other multimedia content like image, video, audio can’t be ignored, hence the concept of keeping the integrity of main content block is better. In this thesis complete main content blocks are filtered and adjusted according to the screen size and displayed on mobile browser.

3. In their work they have mentioned the research gap where user perceived content can be displayed on the mobile web page. This gap has been implemented in this research work and a user perceived view of web page is provided. The user is asked to enter the query in the search page. Using the content matching and similarity of visual block text content with the query terms, the relevant blocks are picked and displayed to the user in the web page. This implementation results in better utilization of screen space and provides light weighted web page for mobile Internet user.