Chapter 1

INTRODUCTION

1.1 Motivation and Objective:

The WWW continues to grow at an explosive rate as an information gateway and as a medium for conducting business. Although the web is growing exponentially, the individual's capacity to read and digest content is essentially fixed. The web is a collection of semi-structured and structured information sources often visualized as a huge and complex dynamic mesh. Therefore the full economic potential of web will not be realized unless enabling technologies are provided to facilitate access to web resources [8][76].

It is also reasonable to believe that data collected over an extended period contains hidden knowledge about business or patterns by characterizing users profile and behavior. Moreover, with the rapid increase of www, the study of knowledge discovery in web, modeling and predicting the users access on a website has become very important for organizations to gain an edge over competitors.

Web servers record and accumulate data about user interactions whenever requests for resources are received. Analyzing the Web access logs can help to understand the user behavior and the web structure. From the business and applications point of view, knowledge obtained from the Web usage patterns could be directly applied to efficiently manage activities related to e-business, e-services, e-education and so on [4][5][56]. Accurate Web usage information could help to attract new customers, retain current customers, improve cross marketing/sales,
effectiveness of promotional campaigns, tracking leaving customers and find the most effective logical structure for their Web space [14][73]. User profiles could be built by combining users’ navigation paths with other data features, such as page viewing time, hyperlink structure, and page content [11]. Several works had been addressed to explain, what makes the discovered knowledge interesting. Results previously known are very often considered as not so interesting. So the key concept to make the discovered knowledge interesting will be its novelty or unexpectedness appearance [1][2][7].

There are many number of commercial software tools that could provide Web usage statistics. These may be useful for Web administrators to get a real sense of the actual load on the server. For small web servers, the usage statistics provided by conventional Web site trackers may be adequate to analyze the usage pattern and trends. But as the size and complexity of the data increases, the statistics provided by existing Web log file analysis tools may prove inadequate. So more intelligent mining techniques are necessary [14].

One of the key steps in knowledge discovery in databases [40] is to create a suitable target data set for the data mining tasks. In Web mining, data could be collected at the server level, client level, proxy level or obtained from organizations databases (which contains business data or consolidated web data). These data could differ in terms of content and the way it is collected etc. The usage data collected at different sources will represent the navigation patterns of different segments of the overall Web traffic, ranging from single user, single site browsing behaviour to multi-user, multi-site access patterns. Web server log does not accurately contain sufficient information for predicting the behaviour at the client side as they are related as they related to the pages given by the Web server.
Log file is a group of all activities which are collected from the web pages irrespective of the error messages with other technical information like session, cookies etc.

Pre-processed and cleaned data could be used for pattern discovery, pattern analysis, Web usage statistics and generating association/sequential rules. Much work has been performed on extracting various pattern information from offline Web logs[39][74] and the application of the discovered knowledge can be used for improving the design and structure of a Web site and helps business organizations to function more efficiently[15,20,25,26,27,28,29,37,58,60,72,73].

To achieve the efficiency and improve the design and structure of a web site by predicting the user's behavior, we have designed a framework which can cover most of the activities of 'web usage mining with a tool known as “Online Miner”[30].

To demonstrate the efficiency of the proposed framework web access log data of “www.primsms.org is used for experimentations.

In Online Miner, several methods are mentioned to extract useful information for proactive personalization of web sites by using Online Miner architecture technology and mining repository algorithm.

Online Miner is an Asp.net technology based framework with C# coding to avoid common problems associated with processing Server Logs and to capture additional and more detailed data. The core elements of this framework are clearly mentioned in the Online Miner architecture.
1.2. Scope and Structure of the Study

Web Mining is a very hot research topic, which combines two of the active research areas: Data Mining and World Wide Web. In general Web usage mining part of Web Mining is the application used in data mining to analyze and discover required hidden patterns of user's usage data on the web. The usage data records the user's interactions when the user browses or makes transactions on the web site. An activity involves the automatic discovery of patterns from one or more Web servers. Organizations often generate and collect large volumes of data; most of this information is usually generated automatically by Web servers and collected in the server log. Analyzing such data can help these organizations to determine the importance/value of particular customers, cross marketing strategies across products and the effectiveness of promotional campaigns, etc.

In this study an attempt has made to achieve the efficiency and improve the design and structure of a website by predicting the users behavior, we have designed a frame work which can cover most of the activities of Web usage Mining with a tool known as "Online Miner".

In chapter 2 all the introductory concepts of Data mining along with its types are discussed. It also includes all the attributes of Web Mining in a detailed manner. In this chapter a brief description of data mining techniques are given. The unstructured feature of Web data triggers more complexity of Web mining. According to Kosala and Blockeel[34], Web Mining research is actually a converging area from several research communities, such as Database, Information Retrieval, Artificial Intelligence and also Psychology and Statistics as well.
Chapter 3 includes the studies such as General Web Usage Mining systems, which aim to discover general trends and patterns from the log files by adapting data mining techniques. The objective of the Site Modification systems is to improve the design of a web site by suggesting modifications in its content and structure. The research on System Improvement focuses on using the web usage mining for improving the web traffic. Finally, personalization systems aim to understand individual trends used for personalizing the web sites.

The application developed and explained in detail in chapter 4 of this thesis. In this Chapter Online Miner features, architecture, process, data flow, results and technological specifications can be found in order to understand the system capabilities easily.

In Chapter 5 presents the code details and program constructs used for designing "Online Miner".

The functionality of Online Miner was explained with a case study and results have been presented in Chapter 6. Benefits of the Online Miner are also explained with inherent illustrations.

Chapter 7 highlights main findings and suggestions for the future work to enhance the system.