CHAPTER - 1

1 INTRODUCTION

As quoted by David Shenk, we are living in a world that is surrounded by data smog [1]. It is nothing but the profuse data which is obtained after accessing the internet. This abundant data however is not able to cater to the growing demands of user. Hence, the most challenging task is converting this vast amount of data into information. The traditional techniques and tools available cannot provide effective results. Data mining combines the power of efficient algorithms with traditional data analysis for converting this raw data into information for analysis [2]. It uncovers the hidden information present in the data and is a major component of Knowledge Discovery. Data mining can be applied in all realms. In context of web, data mining can prove to be highly beneficial. Web mining applies data mining techniques on web resources. The web resources could be in the form of content on web pages, intra and inter page structure, usage data recorded in web access logs or user profile data [3]. By using web mining information extraction from these web resources can be carried out. Web mining can be classified into Web Content, Web Usage and Web Structure mining. Web content mining is applies data mining techniques on web page content. Usage mining is the analysis of web logs for studying the behavior of customer. In Structure mining, web hyperlink structure is examined. The World Wide Web (WWW) is serving as a huge reservoir of data. The web has grown to about 4 billion pages in 2004 and around 8 billion pages by 2006 [4]. Users are facing a lot of difficulty while accessing information on WWW. Hence this plentiful data needs to be mined. An area which looks after this is called web personalization. It looks after the specific interest of the user and suggests which pages to visit next. Web personalization is in great demand these days and is being applied in
almost all the areas. Personalization can be offered by using any form of web mining and combining it with technologies like Collaborative Filtering (CF), Markov Models, Clustering, Association Rule Mining. It could be of any of the three types using manual techniques (user registration preferences), collaborative filtering or content based.

In this thesis, usage and content mining have been utilized for generating personalized suggestions to user. In the proposed work, emphasis has been laid on designing of a tool which can help the users in efficient retrieval of information thereby reducing their effort and time. User profiling has been considered a key feature while designing this personalization service.

1.1 INTRODUCTION TO INTERNET

Internet refers to a network of computers which links various computers anywhere immaterial of the place they reside in. It is a network of networks. It follows a common set of communication protocols. Internet is said to evolve from ARPANET. It was the first Wide Area Network (WAN) having only four sites in 1969. In the beginning it was just used for exchange of information. However in 1989, the restrictions imposed on the usage of internet were removed thereby allowing its usage for commercial purposes\(^5\). Since then internet is becoming predominantly famous amongst all age groups and various sections all over the world.

The growth of internet is increasing every day. As per the survey conducted by Internet in India (I-Cube) 2011 reported internet users in India are constantly increasing. The landmark of 100 million was crossed in September, at 112 million users (88 million urban + 24 million rural villages) who have used the internet even once. Amongst 112 million claimed internet population, 90 million users were ‘active’ internet users i.e. one’s accessing the internet at least once a month.
INTRODUCTION

Internet offers a great deal of services. Typical services include:

- E-mail
- File Transfer Protocol (FTP)
- Usenet
- Telnet

1.2 INTRODUCTION TO DATA MINING

Data mining refers to the collection of exploration techniques based on advanced analytical methods and tools for handling huge data [4]. It is also referred to as Knowledge Discovery in Databases (KDD). It is extracting information hidden in the database. It can be either related to machine learning or exploratory data analysis that is
being widely used these days. Some of the areas include prediction and description, relationship marketing, customer profiling, outlier identification and detecting fraud, customer segmentation, website design and promotion. Basic data mining tasks can be divided into the following:

1. **Classification**: It can be used when we are mapping data into predefined groups or classes. There is a training set wherein every object is represented by a vector of attributes along with its class. Later a model or classification function is built based on the data contained in the training set. This classification function is used for classifying future objects and developing a better understanding of the classes of objects in the database. Classification is considered to be a part of supervised learning. Pattern recognition and Prediction are also covered under classification [6][3].

2. **Regression**: maps a data item to real value prediction variable.

3. **Clustering**: This technique assembles data of similar types in clusters. It is a form of unsupervised learning. The clusters are formed for the group of classes which are unknown and hence it is considered to be a part of unsupervised learning. The major objective of clustering is to minimize the interclass similarity and maximize the intra class similarity depending upon some criteria defined on object of attributes. After deciding upon the number of clusters formed labeling of objects with corresponding clusters is carried out. The common features of objects in the cluster are summarized for forming class description. Segmentation and partition are also a part of clustering.

4. **Summarization**: helps in mapping of data into subsets having associated simple descriptions. It can also be referred to as abstraction or generalization of data. Task relevant data is taken, summarized and abstracted resulting in a smaller set
which provides an overview of the data usually supplemented with aggregate information.

5. Link analysis is used to uncover relationships from the data. Affinity Analysis, Association Rules, Sequential Analysis all come under Link Analysis.

1.3 WEB MINING

Web mining applies techniques of data mining on web data. It uncovers hidden knowledge related to users as well as websites. It helps in understanding customer’s behavior and evaluating the effectiveness of any website. The intrinsic properties of web pose a lot of challenges. Firstly voluminous data is present on the web, secondly data present in web pages are semi structured, and thirdly information on web tends to be diverse in meaning. Web mining techniques if applied successfully can help in resolving all the challenges which have been discussed above. Some other techniques such as databases, information retrieval and machine learning are also being used in order to cope up with these challenges [7]. Web mining is classified into Web usage, Web Content and Web Structure Mining.

1.3.1 Web Usage Mining

This technique helps in understanding users’ behavior by examining his website access patterns. These patterns are recorded in web logs of the particular website. Web logs are generated on the server every time the user clicks the website. Formats like Internet Information Server (IIS), NCSA or custom are followed in recording web logs. Normally, IIS 7 stores data in the format followed by World Wide Web Consortium (W3C). The information recorded in the web log servers can be customized. Fields recorded in the IIS log file format are Client IP address, User name, Date, Time, Server name, Server IP address, Time taken, Client bytes sent, Server bytes sent, Service status
code (A value of 200 indicates that the request was fulfilled successfully.), Windows status code (A value of 0 indicates that the request was fulfilled successfully.), Request type, Target of operation, Parameters (the parameters that are passed to a script). The logs obtained cannot be directly used for analysis and needs to be pre-processed. The following phases are included in web usage mining: Collecting the data, Preprocessing of data, Pattern Recognition and Pattern analysis. Data collection can be carried out from Web and Proxy Servers or from Client Side. Data preprocessing is the second phase and is considered an integral phase in Web usage mining. Irrelevant entries are discarded during this phase, which results in significant reduction of data. During pattern recognition, data mining techniques are applied on the preprocessed data obtained from the last step. Application of relevant mining algorithm is crucial in this context. The last phase i.e. pattern analysis selects appropriate pattern from the available patterns. Visualization techniques can be applied on the extracted patterns for analysis purposes.

Preprocessing of data can be divided into cleaning the data, user identification, session identification and finally path completion. Cleaning data eliminates data with failed HTTP request, data having file extensions .jpeg, .gif, .png etc. and robots.txt files. Data cleaning is a crucial step and results in considerable reduction in data. User Identification refers to tracking the number of users accessing the website. This is accomplished by checking the IP address of the user recorded in the web logs of the website. The major problem being faced here is the IP address conflict. User identification faces another issue wherein if the user agent is different (even if the user is accessing from the same machine), IP addresses returned are different. The concept of cookies can be used as a solution to the above stated problem. Cookies are piece of data sent from website residing in user’s browse while he is accessing a site. Cookies may be referred to as either HTTP cookie or web cookie or Internet cookie, or even browser cookie. During.
our research work cookies were assigned to every user logging onto our website. User who visited the website first time was issued a fresh cookie ID. For every subsequent visit the same cookie ID was utilized. Session identification helps in ascertaining the amount of time spent by every user on a particular website. Session is said to begin as soon as a user logs into a website and terminates when he logs out. Number of web pages a user accesses during a logical period is referred to as a session. The last phase in pre-processing is path completion. Due to proxy servers, caching and pressing browsers back button all the page accesses are not recorded in the web logs [8][9]. For carrying out proper pattern analysis missing pages are appended.

1.3.2 Web Content Mining

Examining content present on web pages along with the results retrieved after surfing the web is referred to as Web Content mining. Web provides a suitable platform for data mining due to these following characteristics [10]: Web offers huge and diverse information, the amount of information present on web is growing everyday at a rapid pace, data of a variety of types exist on web, which could be in the form of unstructured, semi and structured data.

The unstructured refers to the data present in text documents. It is said to be associated with Natural language processing, text mining and machine learning. Web page content and search result are two forms of Web content mining. Structured data could be in form of list, tree or data in tabular format. This type of data could be extracted using Wrapper generation [11]. Unstructured mining is similar to text mining. The techniques included in this are Information Extraction, Topic Tracking, Summarization, Categorization, Clustering and Information Visualization. Techniques employed for extracting structured data are Web Crawler, Wrapper Generation, Page content Mining. For extracting semi
structured data Object Exchange Model (OEM), Top down Extraction, and Web Data Extraction language are some techniques which can be deployed.

### 1.3.3 Web Structure Mining

It is another form of Web Mining used for identifying the relation between web pages. Web Structure mining aims at reducing irrelevant search results and inability to index the vast amount of information provided on the Web. It uses graph theory for analyzing the node and connection structure of website [12]. HITS and Page rank are two well known algorithms being used in Web structure Mining. The major challenge being faced is how to deal with hyperlink structure existing within the web. Structure mining is also associated with Link mining which is an emerging these days in the research industry. It is an instance of multi relational data mining. It includes descriptive and predictive modeling [13]. Web is said to compose a variety of objects like web pages, links etc. and no unifying structure [7]. Hence by using web structure mining strategic results for marketing of a Web site can be obtained. This type of mining is apt if we want to use our website as a business tool.

### 1.4 WEB PERSONALIZATION

The demand for web mining is increasing everyday due to the increase in number of user’s everyday on the web. Reaching the desired location on the web is a major problem that almost every user faces. In order to cope up with this problem the Concept of Web Personalization came into practice. Web Personalization refers to studying the users’ behavior and delivering their custom content that could be helpful to him. It extracts knowledge from users behavior recorded in web logs (Web Usage Mining) and can be easily combined with content (Web Content Mining), structure (Web Structure Mining) as well as user interests and preferences [14]. These days Personalization is attaining
prominence almost in every area. It is indeed helpful for the users as it saves a lot of time. Web personalization is a four step process which starts with data collection, followed by preprocessing, data analysis and at the end interpretations of the results obtained by applying visualization techniques.

The technologies used in Personalization include Collaborative Filtering (CF), Markov Models, Clustering, Association Rule Mining etc. These technologies when collaborated with web data help in generating interesting recommendations to the user. Web personalization can be implemented by using techniques such as Content-based filtering, Collaborative-based filtering or Rule based filtering.

Content-based filtering emphasizes the preferences of every user visiting the website. In this the available resources and user interests are compared and similarity is calculated. The items that are recommended to one are based on users past behavior. Candidate items are compared with items rated previously and best matching items are suggested to the user. User profiling is carried out based on the preferences stated by the user and history of his interaction with the system. Such systems are unable to learn users’ preferences from their actions.

Collaborative filtering systems are based on users’ rating of objects or revealing their preferences and interests and then return information that is predicted to be of interest for them. It assumes that users having similar behavior will have similar interests. Here the emphasis is on information on users’ behaviors is collected and analyzed, activities or preferences and predicting what users will prefer depending upon the kind of similarity they share with others. This technique is not dependent on machine analyzable content it can make accurate recommendations. It just assumes that people will like the items in a more or less similar manner according to their previous tastes. However it suffers from the problem of cold start and sparsity. While user profiling is carried out data collection
can be carried out implicitly or explicitly. Implicit data collection could be done by observing the items that a user has viewed, keeping a track of his purchases etc. On the other hand explicit data collection could be asking users to search or rate an item or may be a collection of item etc.

Rule based filtering is nothing but Association rule mining wherein the items that are brought together in the data set, are looked upon. Items could refer to anything may be a product that is purchased by the customer or set of web pages visited by him. In rule-based filtering if the behavioral pattern of customer matching of the left-hand side of a rule, then the right-hand side can be used for recommendation or prediction. Measures Support and confidence are used for indicating the effectiveness of association rule. Due to numerous rules, returned filtering has to be done so that non-significant associations can be eliminated easily. To achieve this, the concept of minimum support and confidence is used. Frequent pattern growth (FPgrowth) is an important algorithm used in this which is based on divide and conquer strategy.

Apart from utilizing Content-based, rule-based and collaborative filtering techniques independently a combination of any of these approach also known as hybrid, can also help in providing more interesting and accurate recommendations [15].

1.5 E-LEARNING

The usage of electronic technology in teaching and learning has been referred to as E-learning. The extensive use of technology has plagued the Indian society. Almost all age groups are depending highly on the use of internet. The internet is satisfying all the users and hence is generating huge response. Statistics have shown that the use of internet and technology is increasing every day. However the Indian infrastructure is unable to utilize the available resources effectively. If internet and technology is integrated with classrooms and teachers more fruitful results can be obtained. Using the internet
technology will have a global impact and the reach would also increase. Kids spend most of their time accessing tablets, PC’s, Laptops or mobile phones. If knowledge is disseminated from these platforms the students might find it more interesting. It is an extension to classical form of teaching. The e-learning concept ensures that the learning material is available online and also facilitates the mechanism of feedback. Feedback from the students help in ensuring that the course material provided is relevant and is able to satisfy their queries completely. Web based, computer based, virtual classroom and digital collaboration are some applications of E-learning. It increases the power of traditional text book materials when combined with online resources. Students get a good feel sitting at their home looking at rich media and interactive content. On sitting in this virtual classroom one can have freedom of covering the topics which was skipped initially. The student if interested can also extend discussion beyond classroom. Tools required in e-learning are blogs, newsgroups, bulletin boards, web forms, wiki, polling, instant messaging etc. E-learning applications if implemented properly can raise the level of education and literacy. It has been seen that the e-learning has been more successfully implemented in corporate sector. E-Learning offers a lot of advantages. Ease in accessing the content across the globe, 24x7 availability of content thereby, adds onto user’s flexibility, mentoring support. Self pace mechanism are some of the advantages which one can have while the user (student) is accessing the e-learning website.

1.6 PROBLEM DEFINITION

In the current state of art the profuse data present on the web is adding woes to the increasing demands of the users accessing the net. We are considered to be data rich but information poor. On the other hand businesses are facing a lot of challenges in attracting customers to their website. Due care has to be taken while the websites are being designed. Customers want that their queries should be resolved at just one click. They
want they should only get relevant information. Hence to address this changing need of
customer, the concept of web personalization was introduced. Web personalization
customizes the website by analyzing the user’s navigational data (data from web access
logs) and combining it with content, structure or user profile data [16].

During this research, leading web-based application areas such as E-commerce, Internet
Shopping Mall, E-learning, Research Papers where web personalization has been
incorporated, were studied. A comparison of personalized being offered was carried out
based on the following parameters viz. Domain knowledge, degree of personalization,
access time and user profiling.

The above mentioned parameters which were chosen can be considered essential but not
necessarily sufficient for any web-based application. But it can be said that for an
efficient and effective web-based application, the application must adhere to these
parameters. Efficiency aspect is related to how many new customers it can attract.
Effectiveness deals with ability of the system to retain its existing customers. Depending
upon the type of Web Application being designed apart from the parameters suggested a
few other things can be considered while designing Personalization Services like
feedback of user, past study pattern of the learner, whether the user has previously visited
the website, privacy of the customers (thereby increasing customers loyalty). As we are
talking about personalization in web-based applications, the very obvious parameter is
the degree of personalization being offered. It refers to extent the designed system can
help in finding relevant information and the [17] quality of recommendation provided to
the user. Higher the quality of recommendation, higher is the degree of personalization.

For providing the user with relevant information and recommendation, domain
knowledge is the key parameter which must be taken care of while designing the system.
Better domain knowledge helps in attaining the system scalability and increasing its
prediction accuracy [18]. Another vital parameter is User profiling, which takes care of the user’s interest and preferences. User profiling can be judged in a better manner by knowing the user’s hobbies, interest areas [19]. Better the user profiling, better is the degree of personalization. Next factor is the access time. This tells us about the time spent by the user on each web page. This factor is essential to record, because it will tell the interest of the user in which particular web page he is concerned more, which will ultimately help in refining the quality of recommendation and in turn increase the degree of personalization.

Henceforth, the parameters decided will help in ultimately deciding the degree of personalization being offered by the application.

1.7 PROPOSED SOLUTION

Web Personalization offered in several areas of WWW were studied. It was found that areas such as E-commerce, Internet Shopping Mall, E-learning, Research Papers etc. are extensively using this technology but that major work has been done in context of E-commerce. It was found that as compared to E-commerce, personalization was not being used effectively in E-learning. Therefore, for carrying out the research, E-learning website was chosen.

Although a variety of technologies are there which can be used for designing personalization service but majorly CF is being used. However, CF if applied alone can lead to various disadvantages. Not only CF but almost all technologies if used alone can lead to several disadvantages. So a blend of two or hybrid technology can lead to better results and offer higher degree of personalization. User’s feedback, access time criticality, user profiling, domain knowledge are some of the factors which should be taken into consideration while designing Personalization services. Apart from these, few
other parameters can be looked upon depending upon the type of web application for which Personalization Service is being designed.

In the proposed work, user profiling was considered a key feature for designing personalization service. User profile displays a specific user’s personal data. It can also be referred to as representing user model using computer. Profiling is construction of a profile via the extraction from a set of data. The major aim of creating user profiles is collecting user specific interest thereby improving information access quality. This process includes gathering information about user, constructing user profile and finally a technology that can exploit created user profile for providing personalized services.

After creating user profile the user is provided some suggestions depending upon the browsing behavior. The suggested links are available to the user on the second visit i.e. after the user profile has been created. The personalized results are obtained by extracting the content from other websites along with the data present on our own website related to the topics in which the user is interested. A content extraction panel has been inserted which examines the content present on other web pages and depending upon their site rank extracts the links and suggests them to the user. For accessing the performance of the proposed work the mechanism of feedback has also been incorporated. Once the user has accessed the website he is requested to fill in a feedback form. The feedback form has been developed using Google docs. The comments provided by the users are later checked which can be quite useful in satisfying the users.

1.8 THESIS ORGANIZATION

The organization of thesis is as follows:

Chapter 2 talks about the literature work carried out in this area.
Chapter 3 shows the significance of data preprocessing and how it helps in discarding the redundant data and reducing file size. The algorithms required and proposed tool has been presented in this section.

Chapter 4 elaborates on tools and techniques by which data analysis was carried out.

Chapter 5 enumerates the various technologies behind personalization services, the areas in which they are deployed highlighting their advantages and disadvantages. It also gives a brief description of comparison of various personalization services.

Chapter 6 gives an in depth study of the system designed during the research work. It discusses the methodology used in this research along with the user interface of the system.

Chapter 7 presents the results and observations obtained during this research.

Chapter 8 enlists the conclusions on the proposed work and proposes some possible future enhancements.