1. INTRODUCTION

As the World Wide Web is growing rapidly day by day, the numbers of web pages are increasing into millions and trillions around the world. Web is a collection of interconnected documents and other resources, linked by hyperlinks and URLs and it is accessed via the Internet. Search Engines are used to find specific information on the web. With a Web browser, a user views Web pages that may contain text, images, videos, and other multimedia and navigates between them using hyperlinks. This chapter describes overview about World Wide Web and its technology. Web Data Mining and its characteristics are presented. Three knowledge discovery domains in web mining namely Web Content Mining (WCM), Web Structure Mining (WSM) and Web Usage Mining (WUM) are described. Web Content Mining refers to the discovery of useful information from web contents, including text, image, audio, video, etc. Web Structure Mining studies about the hyperlink structure of web. It usually involves analysis of the in-links and out-links of web pages. Web Usage Mining is extraction of the navigation patterns as the browsing patterns could be traced and the structure of the website can be designed accordingly.

1.1 WORLD WIDE WEB

The World Wide Web is defined as a “wide-area hypermedia information retrieval initiative aiming to give universal access to a large universe of documents”. Web is an Internet-based computer network, which allows the users of one computer to access information stored on another computer through the world wide network called the Internet. For the implementation of web, Client-Server technology is used. In this technology, users rely on a program, which is called as Client and it’s
used to connect to a remote machine. The remote machine is called as Server where the data is stored. Navigating through the web is done by a client program called the browser, e.g., Internet Explorer, Mozilla, Alta Vista, Netscape, Firefox, etc. Web browser works by sending requests to remote servers for information and then interpreting the returned documents and laying out the text and graphics on the user’s computer screen on the client side. The operation of the web relies on the structure of its hypertext documents. Hypertext allows web page authors to link their documents to other related documents residing on computers anywhere in the world. To view these documents, one simply follows the links, which is called as hyperlinks [CHM01]. Web has many unique characteristics; some of the characteristics of web are as follow.

I. The huge amount of information is available on the web and is still growing rapidly. The coverage of the information is also very wide and diverse. Hence users can find information on almost anything on the web.

II. Data of all types exist on the web, e.g., structured tables, semi structured web pages, unstructured texts, images, audios, and videos.

III. Web is also about services. Most commercial web sites allow people to perform useful operations at their sites, e.g., to know the current rate of products, to purchase products, to pay bills, and to fill in forms.

IV. Information on the web changes constantly. Keeping up with the change and monitoring the change are important issues for many applications. Hence the web is dynamic.

V. Web is a virtual society. The web is not only about data, information and services, but also about interactions among people, organizations and automated systems. One can communicate with people anywhere in the world
easily and instantly, and also express one’s views on anything in Internet forums, blogs and review sites.

All these characteristics present both challenges and opportunities for mining and discovery of information and knowledge from the web. With the explosive growth of information sources available on the World Wide Web, finding relevant information is becoming more important for web users. With the rapid growth of the web, Web mining is becoming increasingly significant and popular [BLU07] [WWW11] [DHZ99].

1.2 WEB MINING

Emerging field of web mining aims to discover useful information from the web structure, page content, and usage data. Web mining is the application of data mining techniques to extract knowledge from web data, including web documents, hyperlinks between documents, usage logs of web sites, etc. Search engines have a significant role to play in the information retrieval and are used by millions of people all over the world. Web Content Mining (WCM), Web Structure Mining (WSM) and Web Usage Mining (WUM) are the three knowledge discovery domains in web mining [WMW10].

1.2.1 WEB CONTENT MINING

Web Content Mining is the process of mining useful information from the contents of web documents. Content data corresponds to the collection of facts for a web page was designed to convey to the users. It consists of text, images, audio, video, or structured records, such as, lists and tables. Research in web content mining includes resource discovery from the web, clustering such as document based
clustering and information Retrieval from web pages. Application of text mining to web content has been the most widely investigated [BKA00].

1.2.2 WEB STRUCTURE MINING

Jaideep Srivastava et al., 2005, proposed that the structure of a typical web graph consists of web pages as nodes and hyperlinks as edges connecting related pages. Web Structure Mining is the process of discovering structure information from the web. A significant quantity of information on the web is linked. This can be further divided into two kinds based on the kind of structure information used namely Hyperlinks and Document Structure [JPV05].

I. Hyperlinks: A Hyperlink is a structural unit that connects a location on a web page to a different location, either within the same web page or on a different web page. Hyperlinks exist between web pages within a site, and serves as information organization mechanisms. A hyperlink that connects to different parts of the same page is called an Intra-Document Hyperlink. A hyperlink that connects two different pages is called an Inter-Document Hyperlink [HZG00][KKR99]. Hyperlinks exist across different sites, represent implicit conveyance of authority to the target pages. Hyperlinks consists of

a) Link Mining,

b) Identify Authoritative Web Pages

a). Link Mining: The research of structure analysis is called as Link Mining. M.Gomes da Costa Jr. and Z.Gong, 2005, suggest that the web contains a variety of objects with almost no unifying structure, with differences in the authoring style and content much greater than in traditional collections of text documents. The objects in the WWW are web pages, and links are inbound, outbound and co-citation (page i and j are both
linked to by the page \( k \)). Attributes include HTML tags, word appearances and anchor texts [GKM99][GCG05].

\textbf{b). Identify Authoritative Web Pages:} The hyperlinks contain an enormous amount of latent human annotation that can automatically infer the notion of authority. When an author of a web page creates a hyperlink pointing to another web page, this can be considered as the author’s endorsement of the other page. Every hyperlink does not represent the endorsement and some links are created for another purpose, such as, for navigation or for paid advertisement. If a majority of hyperlink is for endorsement, then the collective opinion will still dominate. The endorsement of a given page by different authors on the web indicates the importance of the page and naturally leads to the discovery of \textit{authoritative} web pages [BMP99].

\textbf{II .Document Structure:} Jaideep Srivastava \textit{et al.}, 2005, suggested that the content within a web page can also be organized in a tree-structured format, based on the various tags within the page. Mining efforts have focuses on automatically extracting document object model (DOM) structures out of documents [JPV05].

\textbf{1.2.3 WEB USAGE MINING}

Web Usage Mining is the application of data mining techniques to discover interesting usage patterns from web data, in order to understand and better serve the needs of web-based applications. Chitraa and Antony Selvdoss Davamani, 2010, suggested that the Web Usage Mining refers to the automatic discovery and analysis of patterns in click stream and associated data collected or generated as a result of user interactions with web resources on one or more web sites [CAD10]. The goal is to capture, model, and analyze the behavioral patterns and profiles of users interacting with a web site. The discovered patterns are usually represented as collections of
pages, objects or resources that are frequently accessed by groups of users with common needs or interests [BLU07] [BKA00] [CHM01][MFS13] [WMW10].

1.3 MAJOR OBJECTIVES FOR DOING RESEARCH IN WEB MINING

➢ Internet citizen access an abundance of information from the World-Wide Web, but it becomes increasingly difficult to identify the relevant pieces of information. Major objective of doing research in this area tries to address this problem by applying web mining techniques.

➢ More than 90% of users navigate one or two top most pages and try to get required relevant documents within the navigated pages and ignore the rest of the pages without getting the needed information. Objective of this research is to provide relevant web pages on the top most of web search result.

➢ Network structure of hyperlinked environment can be a rich source of information about the web content. Therefore web structure mining play vital role in web search.

➢ Studying the existing WSM algorithms for possible improvement is considered as yet another objective.

➢ Content of web page can also be an important source of information for the search query terms. Hence web content mining also plays an important role in web search.

➢ Studying the existing WSM algorithms and existing WCM techniques to incorporate hybrid web search method and its architecture for efficient web search is considered as yet another objective.
1.4 ORGANIZATION OF CHAPTERS

This section describes about the arrangement of thesis chapters. Chapter 2 describes literature survey of most widely used Web Content Mining (WCM) techniques and Web Structure Mining (WSM) algorithms. Literature survey of Hybrid Web Search (HWS) methods are also presented.

In chapter 3 the preliminary study and experiments were conducted with the sample web graph to analyze the working functionality of the WSM algorithms.

Chapter 4 presents implementation of PageRank algorithm for analyzing it various characteristics using sample web graph. Four different case studies were considered for this analysis.

Chapter 5 has been proposed a novel and innovative score calculation namely Proportionate Prestige Score (PPS) Method. This proposed PPS method computes the PageRank values using non-uniform transition probability distribution of prestige values of each page.

In chapter 6 the preliminary study and experiments were conducted with the sample web graph to analyze the working functionality of the considered WCM techniques.

Chapter 7 describes a novel method namely Hybrid Web Search (HWS) and its architecture. Working functionality of HWS method has been demonstrated with sample web graph.

Chapter 8 describes result analysis of existing WCM techniques and WSM algorithms and comparison with PPS method using real time datasets. Also proposed Hybrid Web Search (HWS) Method has been substantiated with huge volume of web pages with the different query terms and its results were discussed.
Chapter 9 describes conclusion of thesis. As a future scope, PPS method and HWS method using PPS_VSM can be compared with other web mining algorithms to prove its performance and efficiency of information retrieval.