The web is a rich source of information and persists to increase in size and difficulty. Retrieving the necessary web page on the web, efficiently and effectively, is becoming a challenge aspect now days. On every occasion a user needs to search the relevant pages, the user prefers those relevant pages to be at hand. Relevant web page is one that provides the same topic as the original page but it is not semantically identical to original page. As a matter of fact the Web is unstructured data warehouse, which delivers the mass amount of information and also enlarges the complexity of dealing information from different perspective of knowledge searchers, business analysts and web service providers. Web has grown enormously and the usage of web is unbelievable so it is essential to understand the data structure of web. The mass amount of information becomes very hard for the users to find, extract, filter or evaluate the relevant information. This issue lifts up the attention to the obligation of some technique that can solve these challenges. Web mining can be easily used in this direction to carry out the problem with the help of other areas like Database (DB), Information retrieval (IR), Natural Language Processing (NLP), and Machine Learning etc.

Dealing with these aspects, there are some challenges we should take it into account as follow 1) Web is huge. 2) Web pages are semi structured. 3) Web information stands to be diversity in meaning. 4) Degree of quality of the information extracted. 5) Conclusion of knowledge from information extracted.

Web mining can be broadly divided into three categories:

1. Web Content Mining

2. Web Structure Mining

3. Web Usage Mining.

A. Web Content Mining

Web content mining is the procedure of retrieving the information from the web into more structured forms and indexing the information to retrieve it quickly. It focuses mainly on the structure within a web documents as an inner document level.

It is also quite different from Data mining because Web data are mainly semi-structured and/or unstructured, while Data mining deals primarily with structured data. Web content mining is also different from Text mining because of the semi-structure nature of the Web, while Text mining focuses on unstructured texts. Web content mining thus requires creative applications of Data mining and / or Text mining techniques and also its own unique approaches. In the past few years, there was a rapid expansion of activities in the Web content mining area. This is not surprising because of the phenomenal growth of the Web contents and significant economic benefit of such mining. However, due to the heterogeneity and the lack of structure of Web data, automated discovery of targeted or unexpected knowledge information still present many challenging research problems.

Web content mining could be differentiated from two points of view: Agent-based approach or Database approach. The first approach aims on improving the information finding and filtering. The second approach aims on modeling the data on the Web into more structured form in order to apply standard database querying mechanism and data mining applications to analyze it.

Web Content Mining Problems/Challenges

Data/Information Extraction: Extraction of structured data from Web pages, such as products and search results is a difficult task. Extracting such data allows one to provide services. Two main types of techniques, machine learning and automatic extraction are used to solve this problem.

Web Information Integration and Schema Matching: Although the Web contains a huge amount of data, each web site (or even page) represents similar information differently. Identifying or matching semantically similar data is a very important problem with many practical applications.Opinion extraction from online sources: There are many online opinion sources, e.g., customer reviews of products, forums,

blogs and chat rooms. Mining opinions (especially consumer opinions) is of great importance for marketing intelligence and product benchmarking.

Knowledge synthesis: Concept hierarchies or ontology are useful in many applications. However, generating them manually is very time consuming. A few existing methods that explores the information redundancy of the Web will be presented. The main application is to synthesize and organize the pieces of information on the Web to give the user a coherent picture of the topic domain.

Segmenting Web pages and detecting noise: In many Web applications, one only wants the main content of the Web page without advertisements, navigation links, copyright notices. Automatically segmenting Web page to extract the main content of the pages is interesting problem.

B. Web Structure Mining

Web Structure Mining focuses on analysis of the link structure of the web and one of its purposes is to identify more preferable documents. The different objects are linked in some way. The intuition is that a hyperlink from document A to document B implies that the author of document. A thinks document B contains worthwhile information. Web structure mining helps in discovering similarities between web sites or discovering important sites for a particular topic or discipline or in discovering web communities.

Simply applying the traditional processes and assuming that the events are independent can lead to wrong conclusions. However, the appropriate handling of the links could lead to potential correlations, and then improve the predictive accuracy of the learned models.

The goal of Web structure mining is to generate structural summary about the Web site and Web page. Technically, Web content mining mainly focuses on the structure of inner-document, while Web structure mining tries to discover the link structure of the hyperlinks at the inter-document level. Based on the topology of the hyperlinks, Web structure mining will categorize the Web pages and generate the information, such as the similarity and relationship between different Web sites.

Web structure mining can also have another direction – discovering the structure of Web document itself. This type of structure mining can be used to reveal the structure (schema) of Web pages; this would be good for navigation purpose and make it possible to compare/integrate Web page schemes. This type of structure mining will facilitate introducing database techniques for accessing information in Web pages by providing a reference schema.

C. Web Usage Mining

Web Usage Mining focuses on techniques that could predict the behavior of users while they are interacting with the WWW. Web usage mining, discover user navigation patterns from web data, tries to discovery the useful information from the secondary data derived from the interactions of the users while surfing on the Web. Web usage mining collects the data from Web log records to discover user access patterns of web pages. There are several available research projects and commercial tools that analyze those patterns for different purposes. The insight knowledge could be utilized in personalization, system improvement, site modification, business intelligence and usage characterization.

Web usage mining can be defined also as the application of data mining techniques to discover interesting usage patterns from web usage data, in order to understand and better serve the needs of web-based applications. Usage data captures the identity or origin of web users along with their browsing behavior at a web site. Web usage mining itself can be classified further depending on the kind of usage data.

Web usage mining tries to make sense of the data generated by the Web surfer's sessions or behaviors. While Web-content mining and Web-structure mining utilize real or primary data on the Web;

Web-usage mining mines the secondary data derived from the behavior of users while interacting with the web. This includes data from Web server-access logs, proxy-server logs, browser logs, user profiles, registration data, user sessions or transactions, cookies, bookmark data, and any other data that is derived from a person's interaction with the Web.