DATA MINING TYPES AND TECHNIQUES: A SURVEY

Surbhi Anand*

Rinkle Rani Aggarwal**

ABSTRACT

The quantity of data kept in computer files and databases is growing at an exceptional rate. More refine information from these data is expected. Simple structured or query language queries are no more sufficient to support the increasing demands for information. Due to this increase in the amount of data and requirement of extracting more sophisticated information from that data, mining comes into effect. There are mainly three types of mining: data mining, web mining, and text mining. This paper is a survey paper which will focus on the fundamentals of web mining and describe in detail the process and techniques of web-mining. Web mining is an emerging research topic which combines two of the activated research areas: Data Mining and World Wide Web. As data mining aims at discovering valuable information that is hidden in conventional databases, the Web mining aims at finding and extracting relevant information that is hidden in Web-related data. Web mining is categorized into three areas: Web content mining, Web structure mining, and Web usage mining. Web content mining focuses on the discovery of the useful information from the Web contents. The Web structure mining emphasizes on the discovery of how to model the underlying link structures of the Web. Web usage mining discovers the user's usage pattern and tries to predict the user's behaviours. Web data can be categorized as content, structure, usage, user profile.

*Department of Computer Science & Engineering, Thapar University, Patiala

Assistant Professor,** Department of Computer Science & Engineering, Thapar University, Patiala

1. INTRODUCTION

With the advent of the World Wide Web, the data present on the Web has become a vast source of information. Consequently, this increase in the mass of data has turned researcher's attention towards the use of data mining techniques to extract useful information from that data. Data mining is defined as finding hidden information from the data stored in a database and therefore it has been called exploratory data analysis, data driven discovery, and deductive learning [15]. There are of three types of mining: data mining, web mining, and text mining. Web mining combines the two of the activated research areas i.e. Data Mining and World Wide Web. Therefore, Web mining can be defined as the application of data mining techniques in order to discover patterns from the Web data, comprising Web documents, hyperlinks between documents, and usage logs of the websites [7].

World Wide Web is one of the most interactive and popular medium to spread the information today. Data on the web is rapidly growing day by day. In addition web data is gigantic, diverse and dynamic in nature due to which users could encounter the following problems while interacting with the web [8]:

1. **Finding Relevant Information**- Users use the search service to find out the specific information on the web. Today's search tools have certain problems like low accuracy due to insignificance of many of the search results. This results in a difficulty in finding the relevant information.
2. **Creating new knowledge out of the information available on the web**- This problem is data triggered. It assumes that there already exists a collection of web data and need is to potentially use full knowledge out of that data.
3. **Personalization of information**- Users of internet differs in their experience intended for the contents they search for and the presentation of their search result. This lead to the problem of personalization.
4. **Learning about Consumers or individual users**-This problem is about what the customers requires out of the information provided on the web. It includes problems such as customization of the information according to the intended consumers and its personalization to an individual user.

Web mining techniques can be used to solve these problems. Web mining never works in isolation. It is a multidisciplinary field that relates to several research communities such as databases (DB), Information Retrieval (IR), and Natural Language Processing (NLP), artificial intelligence (AI) etc, to solve the above stated problem [8].

Web mining can be categorized into [4] [20] content mining, structure mining and usage mining depending upon which part of the web to mine.

2. LITERATURE SURVEY

2.1 Overview

In 1996 its Etzioni [16] was the first person who coined the term Web mining. Etzioni started the hypothesis that information on web is sufficiently structured and outlines the subtasks of web mining. According to him, Web mining is the use of data mining techniques to automatically discover and extract information from the data stored within the World Wide Web documents.

Therefore, Web mining can be decomposed into following subtasks [8] [16] as described in the figure 1:

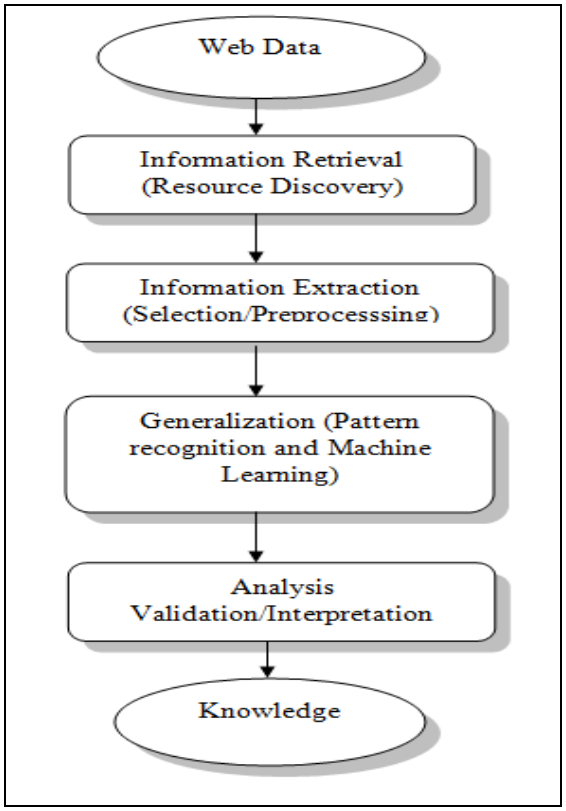

Figure 1 Web Mining Subtask [7]

a) **Information Retrieval (Resource Discovery):** It is the task of retrieving the intended information from the Web. It locates the unfamiliar documents and services on the Web.

b) **Information Extraction (Selection/ Preprocessing):** It is the task of automatically selecting and pre-processing the specific information from the retrieved Web resources.

c) **Generalization (Pattern Recognition and Machine Learning):** It is the task of automatically discovering the general patterns from both the individual Web sites as well as across multiple sites.

d) **Analysis (Validation/Interpretation):** It is the task of analyzing and validating the mined pattern.

Based on the above mentioned subtasks (Figure. 1), web mining can be viewed as the use of data mining techniques to automatically retrieve, extract and evaluate information for knowledge discovery from web documents and services [8][16].

2.2 Web Data

The most essential step in the Knowledge Discovery process is to create a suitable target set of dataset for the mining tasks. In Web Mining data can be collected at the server side, client-side, proxy servers, or can be obtained from an organization's database. The type of data collection differs not only in terms of the location of the data source, but also the kinds of data available, the segment of population from which the data was collected, and its method of implementation [11].

The data that can be used in Web Mining is classified as [11]:

a) **Content:** It is the data that a Web page is designed to convey to the users. This usually consists of text, graphics, audio and video clips, tables, etc.

b) **Structure:** It is the data that describes the organization of the content. It can be of two types: Intra-page structure information and Inter-page structure information. The intra-page structure information includes the arrangement of various HTML or XML tags within a given page. This can be represented as a tree structure, where the (html) tag becomes the root of the tree. The inter-page structure consists of hyper-links connecting one page to another.

c) **Usage:** It is the data that describes the pattern of usage of Web pages, such as IP addresses, page references, and the date and time of accesses.

d) **User Profile**: It is the data that provides demographic information about users of the Web site. This includes registration data and customer profile information.

2.3 Taxonomy of Web Mining

Web Mining can be broadly divided into three distinct categories, according to the kinds of data to be mined [4] [20]:

a) Web Content Mining

b) Web Structure Mining

c) Web Usage Mining

A figure depicting the taxonomy is shown in Figure 2.

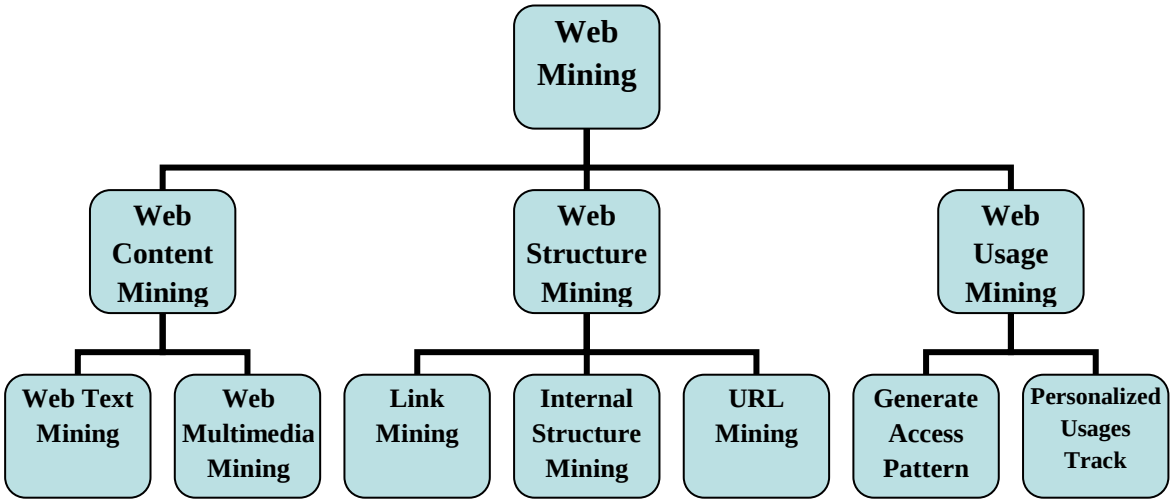

Figure 2 Web Mining Taxonomy [4]

A brief overview of the above three categories [4] [20] is given as follows.

2.3.1 Web Content Mining

Web Content Mining is the process of extracting useful information from the contents of Web documents. Web content mining describes the automatic search of information resource available online [6], and involves mining the contents of web data. In the Web Mining domain, extracting the productive knowledge from the unstructured data residing within the Web documents is analogous to applying the data mining techniques to relational databases. A Web document usually contains several types of data such as text, image, audio, video, metadata and hyperlinks. It can be semi-structured such as HTML documents or a more structured data like the data in the tables or database generated HTML pages, but most of the data is unstructured text data. The unstructured attribute of Web data makes the Web content mining process to be a more complex approach.

The Web content mining is differentiated from two different points of view [3]: Information Retrieval View and Database View. The research works for unstructured data and semi-structured data has been done from the information retrieval view [8]. For the unstructured data, the researchers use a bag of words. In this approach words are isolated from each other to represent unstructured text. Then a single word, which is found in the training corpus, is taken as feature for retrieving the information. On other hand, for the semi-structured data, either the HTML structures inside the documents or the hyperlink structure between the documents are utilized for document representation. As for the database view, the mining process tries to comprehend the structure of the Web site in order to transform a Web site to

become a database. This helps in improved information management and querying on the Web.

Multimedia mining is a part of the content mining, which is engaged to mine the multifaceted information and knowledge from large online multimedia data sources. Multimedia data mining on the Web has recently gained many researchers' attention. Web multimedia mining relates to research areas from several disciplines such as computer vision, multimedia processing, multimedia retrieval, data mining, machine learning, database and artificial intelligence [17]. Text mining is the data analysis of text resources so that novel and previously unknown knowledge can be discovered from the text documents [14]. Text documents differs from the information stored in database systems as they are unstructured by their nature [1].The field of text mining has received plenty of attention owing to the necessity of managing the information that resides in the immense amount of available text documents.

2.3.2 Web Structure Mining

Web Structure Mining focuses on discovering and modelling the link structure of websites [4]. There are many Web information retrieval tools available that ignores the link information of the web pages that could be beneficiary for the researchers and the users. The goal of Web structure mining is to generate the structural summary of the Web sites and Web pages. Web content mining mainly focuses on the structure of the intra-document, while Web structure mining tries to discover the link structure of the hyperlinks at the inter-document level. Web structure mining categorizes the Web pages and generates the information from the topology of the hyperlinks between the web pages to illustrate the similarities and relationships between different Web sites. This inter-document level structure mining can be used to reveal the structure (schema) of Web pages. It solves the purpose of the navigation and helps in comparison and integration of the Web page schemas.

When a Web page is linked to another Web page directly, or the Web pages are neighbours, it is desirable to discover the relationships among those Web pages. These web pages may either be related by synonyms or ontology [20]. When the web pages are synonyms it means they have similar contents. When the relationship between the web pages is ontology it means that they reside in the Web server created by the same person.

Two algorithms that have been proposed for web structure mining are: HITS [10] and PageRank[13]. Hyperlink-Induced Topic Search (HITS) is a link analysis algorithm that rates Web pages. HITS introduce the concepts of hubs (i.e. pages that refer to many other pages) and authorities (i.e. pages that are referred by many other pages). The idea behind Hubs and

Authorities originated from the creation of web pages. During the design of web pages certain web pages were stated as hubs. These pages did not carry any authoritative information, but were used to serve as large directories of information that led users directly to other authoritative pages. In other words, a good hub represented a page that pointed to many other pages, and a good authority represented a page that was linked by many different hubs [5]. Hubs and Authorities can be viewed as 'fans' and 'centers' in a bipartite core of a Web graph [12]. The nodes on the left represent the hubs and the nodes on the right represent the authorities as shown in the figure 4.

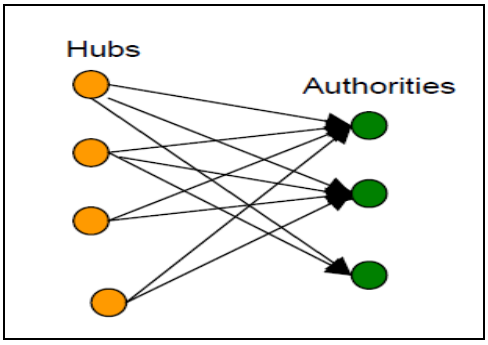

Figure 3 Bipartite Core [12]

Figure 5 shows an example of good hubs and good authorities for web sites of car manufacturers.

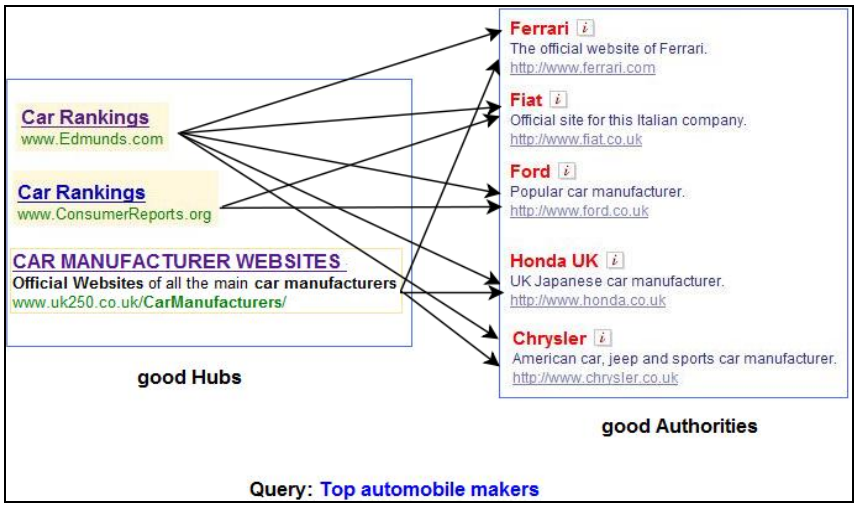

Figure 4 Example showing good hubs and good authorities [21]

PageRank is a link analysis algorithm used by Google search engine that assigns a numerical weighting to each element of a hyperlinked set of documents, such as the World Wide Web, with the purpose of measuring its relative importance within the set [22]. The numerical weight that it assigns to any given element E is referred to as the PageRank of E and denoted by PR (E). A PageRank results from a mathematical algorithm which is based on

the webgraph, created by World Wide Web. The Web can be viewed as a directed graph whose nodes are the documents and the edges are the hyperlinks between them, as shown in the figure 5. The graph structure of the World Wide Web can be used for analysis to improve the retrieval performance and classification accuracy [9]. The rank value indicates an importance of a particular page. A hyperlink to a page counts as a vote of support. The PageRank of a page is defined recursively and depends on the number and PageRank metric of all pages that link to it (i.e. incoming links). A page that is linked to by many pages with high PageRank receives a high rank itself. If there are no links to a web page there is no support for that page [22].

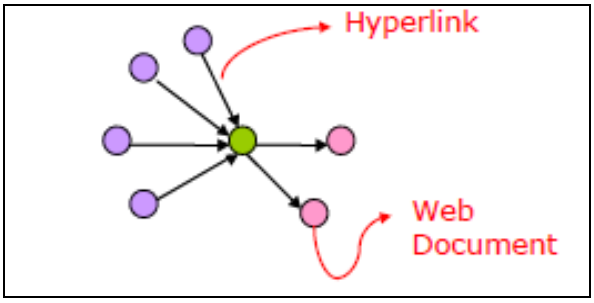

Figure 5 Web Graph Structure [12]

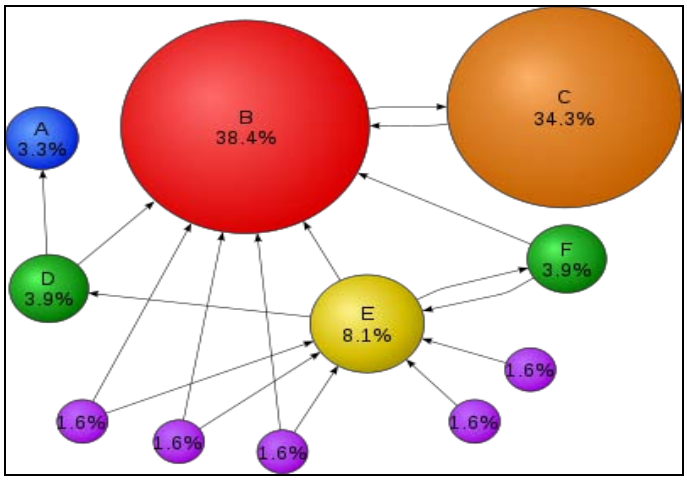

Figure 6 PageRanks (out of 100) for a simple network [22]

The figure 6 shows the PageRanks (out of 100) for a simple network [22]. It shows that the page C has a higher PageRank than page E, even though it has fewer links to it, but the link it has is of much higher value. Page A is assumed to link to all pages in the web, because it has no outgoing links.

Web Structure mining is useful for extracting information such as quality of Web Page in terms of the authority of a page on a topic and ranking of web pages, interesting web structures, web page classification, finding related pages, etc.

2.3.3 Web Usage Mining

A Web server records and accumulates data about the users' interactions whenever user sends request for resources to the web server. This piece of information is named as web logs. The analyzes of the web access logs from different web sites can help to understand the user behaviour and the web structure, thereby improving the design of collection of web resources. Web Usage mining process mines data from log records of web pages. Log records useful information such as URL, IP address and time and so on [18].

The logs of web access available on most of the servers are good examples of the data sets that can be used in web-usage mining. Other web usage data may include browser logs, user profiles, registration files, user sessions and transactions, user queries, bookmark folders, etc. A Web server log is an important source of data on which Web Usage mining can be performed because it keeps track of the browsing behaviour of website visitors.

A Web Usage mining process consists of three phases [11] [19]:

i. Preprocessing,
ii. Pattern Discovery, and
iii. Pattern Analysis

The web usage process is shown in figure 7 below.

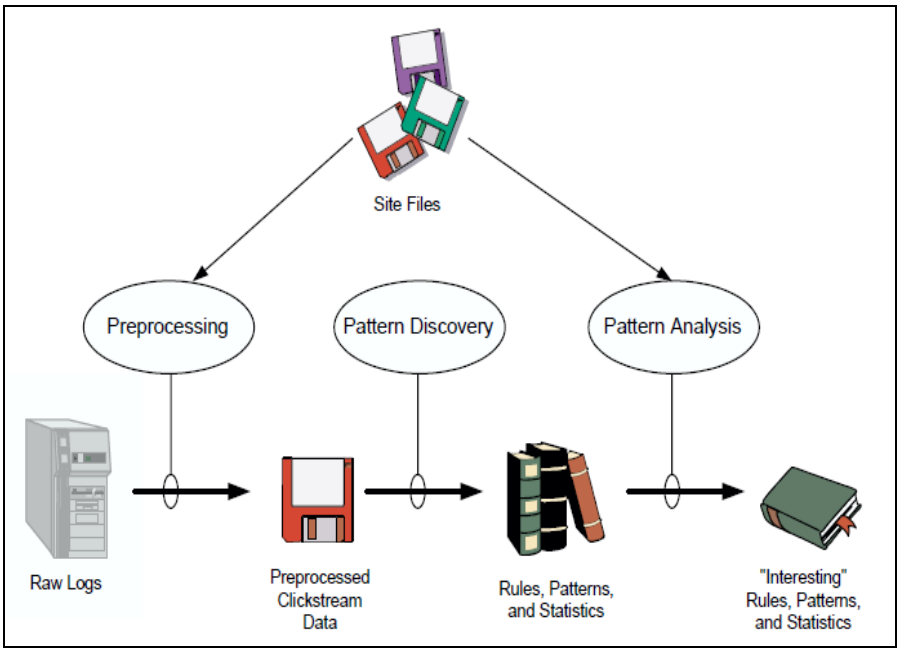

Figure 7 High Level Usage Mining Process [11]

i. **Preprocessing:** Preprocessing consists of converting the usage, content, and structure information contained in the various available data sources into the data abstractions necessary for pattern discovery [11]. Before applying the data mining algorithm, data

preparation, to convert the raw data into the data abstraction, is necessary to perform before moving on to the further processing. The data can be collected at the server-side, client-side, proxy servers, or obtained from database. The data collected, may differ not only in the location of data, but also on the available data type, the segment from where the data was collected and the method of implementation. The information sources available include the Web usage logs, Web page descriptions, Web site topology, user registries, and questionnaire. It has three different conversions: Usage preprocessing, Content preprocessing, and Structure preprocessing [11].

a) **Usage Preprocessing:** Usage preprocessing is the most difficult task in the Web Usage Mining process due to the incompleteness of the available data. The goal of usage preprocessing is to end up with a set of minable objects for a particular Web site [19]. The most common type of input is a Web Server log that can be either in the CLF or ECLF format.

   There are three major required steps and one optimal step to convert raw usage data into server sessions- data cleaning, user/sessions identification, page view identification, and path completion [19]. Data Cleaning is a site specific step that involves merging of logs from multiple servers and parsing the log into data fields. It involves removing the graphics file requests. At this stage the automated agents and spider programs requests are also removed. The final step of this stage is to normalize the URLs. The session identification divides the page accesses of each user, who is likely to visit the Web site more than once, into individual sessions. Another problem is of path completion. It indicates whether there are any important accesses missed in the access log. Finally formatting is performed, which is a preparation module to properly format the sessions or transactions [11] [19].

b) **Content Preprocessing:** Content preprocessing consists of converting the text, image, scripts, and other files such as multimedia into forms that are useful for the Web Usage Mining process [11]. The content of static page views can be easily preprocessed by parsing the HTML and reformatting the information or running additional algorithm as desired [19]. But it is complicated to preprocess the content of dynamic page views.

c) **Structure preprocessing**: The structure of a Website is formed by the hyperlinks between page views. The structure preprocessing can be treated similar to the content

preprocessing. Again, dynamic content poses more problems than static page views. A different site structure may have to be constructed for each server session [11].

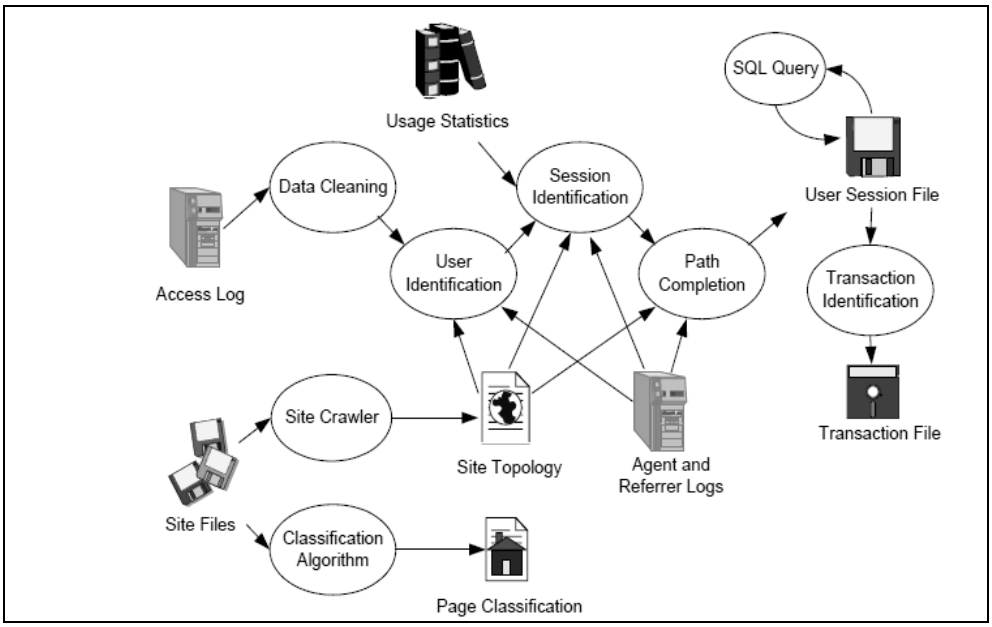

Figure 8 Details of Web Usage Mining Preprocessing [2]

ii. **Pattern Discovery:** This step is the key component of the Web mining. Pattern discovery utilizes the algorithms and techniques from several research areas, such as data mining, machine learning, statistics, and pattern recognition. Some methods of pattern discovery [11]:

a) **Statistical Analysis:** Statistical techniques are the most common method to extract knowledge about visitors of a Web site. By analyzing the session file, one can perform different kinds of descriptive statistical analyses (frequency, mean, median, etc.) on variables such as page views, viewing time and length of a navigational path [11].

b) **Association Rules:** In the context of Web Usage Mining, association rules refer to sets of pages that are accessed together with a support value exceeding some specified threshold. These pages may not be directly connected to one another via hyperlinks.

c) **Clustering:** Clustering analysis is a technique to group together users or pages with the similar characteristics. In the Web Usage domain, there are two kinds of interesting clusters to be discovered: usage clusters and page clusters. Clustering of users tends to establish groups of users exhibiting similar browsing patterns while clustering of pages will discover groups of pages having related content.

d) **Classification:** Classification maps a data item into one of several predefined classes. In the Web domain, this technique is used to establish a profile of users belonging to a particular class or category. This requires extraction and selection of features that best describe the properties of a given class or category.

e) **Sequential Pattern:** It intends to find the inter-session pattern, such that a set of the items follows the presence of another in a time-ordered set of sessions or episodes.

f) **Dependency Modelling:** It tries to create a model that represents significant dependencies among various variables found in the Web domain.

iii. **Pattern Analysis:** Pattern Analysis is the final stage of the Web usage mining. The goal of pattern analysis is to eliminate the irrelevant rules or patterns from the output of the pattern discovery process. There are two most common approaches for the pattern analysis: SQL query mechanism and constructing multi-dimensional data cube to perform OLAP operations [11].

3. CONCLUSION

Web data mining is a fast growing research area today. Extensive literature has been reviewed based on three types of web mining, namely web content mining, web usage mining, and web structure mining. Web content mining focuses on discovering useful information or knowledge from web page contents, while the Web structure mining deals with discovering and modelling the link structure of Web. The main focus of web structure mining is on link information. Web usage mining focuses on understanding user behaviour as depicted in the web access logs while interacting with a website. Web usage mining aims to obtain information that may assist in web site reorganization, website personalization and site adaptation to better suit the user.

REFERENCES

[1] A. Stavrianou, P. Andritsos, N. Nicoloyannis, "Overview and Semantic Issues of Text Mining," Newsletter: ACM SIGMOD Record, Vol. 36 No. 3, pp. 23-34, Sept. 2007.

[2] B. Mobasher, J. Srivastava, R. Cooley," Data preparation for mining world wide Web browsing patterns," Knowledge and Information Systems," Vol. 1 No. 1, pp.5-32, 1999.

[3] B. Mobasher, J. Srivastava, R. Cooley, "Web mining: Information and pattern discovery on the World Wide Web," in the Proceedings of the Ninth IEEE International Conference on Tools with Artificial Intelligence (ICTAI'97), pp.558-567, Nov. 1997.

[4] B. Singh, H.K. Singh, “Web data Mining Research”, in the Proceedings of 2010 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), pp. 1-10, Dec. 2010.

[5] C. D. Manning, H. Schütze, P. Raghavan, "Introduction to Information Retrieval," Cambridge University Press, Cambridge, England, 2008.

[6] E.P.Lim, S.K.Madria, S.S.Bhowmick, W.K.Ng, “Research issues in Web data mining”,in the Proceedings of Data Warehousing and Knowledge Discovery, First International Conference, pp. 303-312, 1999.

[7] G. Srivastava, K. Sharma, V. Kumar," Web Mining: Today and Tomorrow", in the Proceedings of 2011 3rd International Conference on Electronics Computer Technology (ICECT), pp.399-403, April 2011.

[8] H. Blockeel, R. Kosala, “Web mining research: A survey,” ACM SIGKDD Explorations, Vol. 2 No. 1, pp. 1-15, June 2000.

[9] J. Furnkranz, "Web structure mining-Exploiting the graph structure of the worldwide web", in OGAI Journal, Vol. 21 No. 2, pp. 17-26, 2002.

[10] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," in the Proceedings of ACM-SIAM Symposium on Discrete Algorithms, pp. 668-677, 1998.

[11] J. Srivastava, R. Cooley, “Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data”, ACM SIGKDD Explorations, Vol.1 No. 2, pp.12-23, Jan 2000.

[12] J. Srivastava, "Web Mining: Accomplishments & Future Directions", National Science Foundation Workshop on Next Generation Data Mining NGDM02, pp. 1-148, 2002.

[13] L. Page, R. Motwani, S. Brin , T. Winograd."The Pagerank citation ranking: Bring order to the Web," Stanford University, Dept. of Computer Science, Technical report SIDL-WP-1999-0120, Jan.1998.

[14] M. A. Hearst, 1999, "Untangling text data mining," in the Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics, pp.3-10, June 1999.

[15] M. H. Dunham, “Data Mining Introductory & Advanced Topics”, Prentice Hall, 2003.

[16] O. Etzioni, "The world wide Web: Quagmire or gold mine.” Communications of the ACM, Vol. 39 No. 11, pp. 65-68, Nov. 1996.

[17] P. M. Kamde1, Dr. S. P. Algur, " A Survey on Web Multimedia Mining," The International Journal of Multimedia & Its Applications (IJMA), Vol.3 No.3, pp. 72-84, Aug. 2011.

[18] Q. Han, X. Gao, W.WU, "Study on Web Mining Algorithm Based on Usage Mining", In the Proceedings of 9th International Conference on Computer-Aided Industrial Design and Conceptual Design, pp.1121-1124, Nov. 2008.

[19] R. Cooley. Web Usage Mining: Discovery and Application of Interesting Patterns from Web data," PhD thesis, University of Minnesota, Dept. of Computer Science, May 2000.

[20] Y. Wang," Web Mining and Knowledge Discovery of Usage Patterns", CS748T Project Part I, Vol. 7 No. 1, pp. 1-25, Feb. 2000.

[21] "HITS Algorithm - Hubs and Authorities on the Internet", Available:http://www.math.cornell.edu/~mec/Winter2009/RalucaRemus/Lecture4/lecture4.html

[22] "PageRank", Available: http://en.wikipedia.org/wiki/PageRank.