IJCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Volume-6, Special Issue-2, March 2018 E-ISSN: 2347-2693

# Research Issues on Web Mining

**P.Joseph Charles[1*], I.Carol [2], Barna Bass[3]**

[1] Department of information technology, St.Joseph's College, tirchy-02
[2] Department of information technology, St.Joseph's College, tirchy-02
[3] Department of information technology, St.Joseph's College, tirchy-02

**Available online at: www.ijcseonline.org**

***Abstract***— A collection of inter-related files on one or more web servers is known as Web, while web mining means extracting valuable information from web databases. Web mining is one of the data mining domains where data mining techniques are used for extracting information from the web servers. The web data includes web pages, web links, objects on the web and web logs. Web mining is used to understand the customer behavior, evaluate a particular website based on the information, which is stored in web log files. Web mining is generated by making use of the data mining techniques, classification, clustering, and association rules. The collection of information becomes very hard to find, extract, filter or evaluate the relevant information for the users. With the flood of information on the Web, Web mining is a new research issue, which draws great interest from many communities. Currently, there is no agreement about Web mining yet. It needs more discussion among researchers in order to define what it is exactly. In this paper, we have studied the basic concepts of web, web mining, classification, processes, the taxonomy and the function of Web mining.

***Keywords***- Web Mining, Taxonomy, web structure mining

## I. INTRODUCTION

Web mining is the application of data mining technique, which is an unstructured or semi structured data and it automatically discovers and extracts potentially useful and previously unknown information or knowledge from the web [1]. Web mining has three classifications namely, web content mining, web structure mining and web usage mining. The Main web mining Applications are website design, web search, search engines, information retrieval, network management, Ecommerce, business and artificial intelligence, web market places and web communities.

Data mining is used to identify valid, novel, potentially useful and ultimately understandable pattern from data collection in database community [2]. Currently, there is no agreement about Web mining yet. It needs more discussion among researchers in order to define what it is exactly. Meanwhile, the development of Web mining System will promote its research in turn. The World Wide Web (Web) is a popular and interactive medium to disseminate information today. The Web is huge, diverse, and dynamic and thus raises the scalability, multimedia data, and temporal issues respectively. Due to those situations, we are currently drowning in information and facing information overload [3]. Online business breaks the barrier of time and space as compared to the physical office business. Big companies around the world are realizing that e-commerce is not just buying and selling over Internet, rather it improves the efficiency to compete with other giants in the market. This application includes the temporal issues for the users. There are three classifications in Web Mining, which are called as Web content mining, Web structure mining and Web usage mining. The each type of web mining classification consist of its own algorithms and tools to perform in the web.

Web content mining is nothing but the discovery of valuable information from web documents and these web documents may contain text, image, hyperlinks, metadata and structured records. It is the process of retrieving the useful information from the web content or web documents.

Web structure mining is also a process of discovering structured information from the websites. It is used to find the structured data from the existing database of the website. The structure of a graph consists of web pages and hyperlinks where the web pages are considered as nodes and the hyperlinks are edges and these are connecting between related pages.[4] Web usage mining is also called as web log mining. It reflects the user's behavior, which can catch the meaningful patterns from one or more web localities [5].

Web mining process consists of four important steps, they are, resource finding, data selection and pre-processing, generalization and analysis [6]. Resource finding is the either process, which is used to extract the data from online, or offline text resources. In data selection and preprocessing step, specific information from retrieved web sources are automatically selected and pre-processed. During generalization, data mining and machine learning techniques are used to discover general patterns from individual web

sites as well as across multiple sites. Validation and interpretation of the mined patterns are done in analysis step. [7][8]. Web mining is classified into three different categories, they are, web content mining, web structure mining and web usage mining. This is illustrated in Figure 1.


Figure 1. Classification of Web Mining

## II. RELATED WORK

As the web is highly dynamic in Internet; the lots of pages are added, updated and removed every day and it handles huge set of data and information, hence there is an arrival of many number of problems or issues due to such process in web every day. Generally, A web data is high dimensional, which is limited query interface, and a keyword oriented search and a limited customization to individual users. Due to this, it is very difficult to find the relevant and related information from the web, which may create new, and many issues. In Web mining techniques are clustering, classification and association rules which are used to understand the customer behavior and activity to evaluate a particular website by using traditional data mining parameters. Web mining process is divided into four steps; they are resource finding, data selection and pre-processing, generalization and analysis [10] [9]. The web measurement or web analytics in web mining are one of the significant challenges in web mining. The major factors that are used to measure the web analytics are hits, page views, visits or user sessions and find the unique individual visitor regularly used to measure the user impact of various proposed changes. Large institutions and organizations archive usage data from the web sites [11]. Maintaining accuracy in classifying the data needs to be concentrated.

The Important problem is that, identifying, detecting and/or preventing fraud activities and behaviors of the third party. The web usage mining algorithms are more efficient and accurate. However, there is a challenge that has to be taken into Consideration. Web cleaning is the most important process but data cleaning becomes difficult when it comes to heterogeneous data [12].

### *A. Major Problems in Web Mining*

- The Data set in the Web are very larger, and the Storage capacity of the files to be stores in database are around ten to hundreds of terabytes.
- The dataset of individual server cannot be mined, it need large number of servers.
- Limited coverages, limited queries interface to individual users.
- The data cleaning are automated in the process.
- Huge number of over and under fitting of data.
- Huge number of over sampling of data.
- Scaling u for high dimensional data.
- Time series data and timing sequence.
- Very difficult to find the relevant data.
- Extracting new and recent files from the Web.

## III. METHODOLOGY

Taxonomy of Web mining, i.e., Web content mining, web structure mining and Web Usage Mining. The diversity of information on the Web leads to the variety of Web mining, as shown in figure 2. According to the type of source, Web mining can be roughly divided into two domains: Web content mining and Web structure mining. The former is the process of extracting knowledge from the content of Web documents, while the latter is the process of inferring knowledge from the organization and links on the Web.

### 3.1 Web Content Mining

The lack of structure that permeates the information sources on the World Wide Web makes automated discovery of the Web-Based information difficult. Traditional search engines such as Lycos, Alta, Vista, WebCrawler, ALIWEB. Web content mining can be divided into text mining (including text file, HTML document, etc.) multimedia mining.

Web content mining is the mining, extraction and integration of useful data, information and knowledge from Web page content. The heterogeneity and the lack of structure that permits much of the ever-expanding information sources on the World Wide Web, In recent years these factors have prompted researchers to develop more intelligent tools for information retrieval, such as intelligent web agents, as well

as to extend database and data mining techniques to provide a higher level of organization for semi-structured data available on the web. The agent-based approach to web mining involves the development of sophisticated AI systems that can act autonomously or semi-autonomously on behalf of a particular user, to discover and organize web-based information.

The data in Web content mining may be structured or unstructured/semi structured even though many of the web is unstructured. It is the process of retrieving the information from the web into more
Structured forms and indexing the information to retrieve quickly or finding valuable information From web content or web documents. Web content mining includes the web documents, which may consist of text, html, multimedia documents i.e., images, audio, video and sound etc. The search result mining contains the web search results. It may be a structure documents or unstructured documents[*].

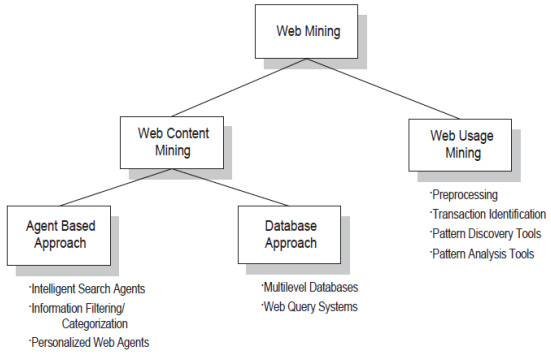



Figure 2. A Taxonomy of Web Mining

Many algorithms and tools are used in Web Content Mining such as Correlation algorithm. Web Info Extractor (WIE), Genetic algorithm, Cluster Hierarchy Construction Algorithm (CHCA), ontology based tools; content mining tools are the web content extractor and automation anywhere. It need to require extracting the information from the cloud provided by web servers by Cloud users, due to which the web servers can make use of the web mining. Web based communities can be maintained the information such as Facebook. The users of same field of interest can be combined or grouped and they can communicate and exchange the information's through the network. Automated citation indexing using web mining techniques are performed by Digital library. E-services include e banking, search engines, online auctions, on-line knowledge management, social networking, e-learning, blog analysis, and personalization and recommendation systems. This can be analyzed for the customers and enable provision to the customers based on their recommendations [13].
It has two types of approaches; they are

(i) Agent based and
(ii) Database Approach.

Figure 3 gives the web content mining approaches.

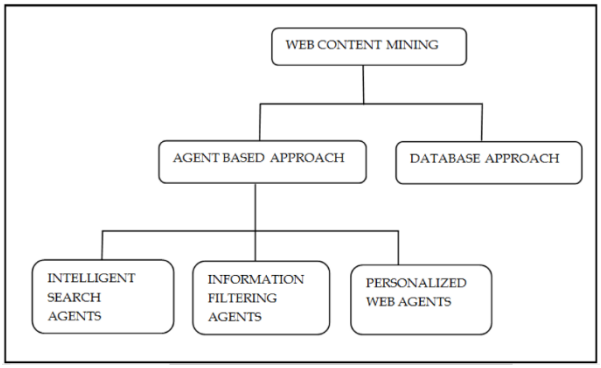



Figure 3. Web Content Mining approaches

**A. Agent Based Approach** -Agent based approach aims on searching and finding relevant information from the World Wide Web (www). Three types of agents they are

(i) *Intelligent search agents* – which means that the system automatically searches for information along with a particular query that are provided by the user.

(ii) *Information filtering/categorizing agents* - Filters the data or extract the data of information from the database.

(iii) *Personalized web agents* – Discovers and finds the documents and files those are related to the user profiles.

**B. Database Approach**

Databases includes the Database approach, which includes tables, attributes and schema with defined and stated domains. It targets on techniques for organizing the semi structured data on the web into more collections of resources, and using standard database querying mechanism and data mining techniques to analyze it, for example multilevel database and web querying system [15]. To mine the data, the Web content mining has the other approaches. These are unstructured text data mining, structure mining, and semi-structure text mining and multimedia data mining [16].

### 3.2 . RESEARCH PROBLEMS ON WEB CONTENT MINING

As we discussed Web content mining can extract the information from the web search engines, which becomes the number of research issues in it.

• *Opinion extraction from online sources* i.e. customer makes sure of products, forums, blogs and chat and living rooms. Big consequence for marketing intelligence and product benchmarking. Are in the process Mining opinions.

• *Data / Information Extraction* concentrate on extraction of structured data from web pages such as products and search results.

• *Automatically segmenting* web pages and identifying or detecting noise and unwanted data is an interesting problem in web application.

• Web information integration and schema matching. The web contains large amount of data, each website accept similar information in a different way.

## IV. WEB STRUCTURE MINING

The study of data interconnected to the structure of a particular website is called the Web structure mining.
Graph theory to analyze the node and connection structure of a web site is used by Web structure mining. The aim for structure mining is to extract previously unknown relationships between Web pages of same website. The mining technique is performed either at the intra-page (document) level or at the inter-page (hyperlink) level.The goal of the Web Structure Mining is to generate the structural abstract about the Web site and Web page. Web Structure mining will categorize the Web pages and provide the information like similarity and relationship between different Web sites.

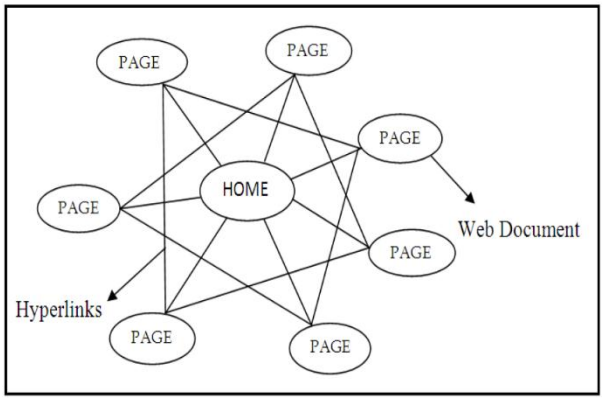



Figure 4. Web Graph Structure

Extracting information is useful source for Web structure. Web structure is to extract some interesting web graph patterns like co-citation, social choice, complete bipartite graphs, etc. [7]. It classifies the web page on various topics and deciding which web page is to be added into the collection of web pages. Web structure mining can be performed either at intra-page level or at inter-page level. A hyperlink that connects to a different part of the same page is called intra-page hyperlink. It is a document structure level [17]. A hyperlink that connects two different pages are called inter-page hyperlink, which is structure level [12]. Some of the important tasks of link mining are link-based classification, link based cluster analysis, link type, link strength and link cardinality. The research of the hyperlink level is also called hyperlink analysis [17], which can be used to retrieve useful information from the web [14].

Web structure mining is used in search engines such as Google, Yahoo, etc. IBM used HITS algorithm in clever search engine and Google [10] uses the page rank algorithm. Algorithms of web structure mining are HITS (Hypertext Induced Topic Search) algorithm, Max flow- Min cut algorithm, ECLAT algorithm, and Page rank algorithm. Page rank algorithm can be divided into two types. One is weighted page rank algorithm and another one is Topic sensitive page rank algorithm.

### 4.1 RESEARCH ISSUES ON WEB STRUCTURE MINING

Web structure mining has two issues due to its huge amount of data in database.

(a) Reducing irrelevant search results. Relevance of search information becomes unorganized due to the problem search engines often only tolerate for low precision criteria.
(b) Indexing information on the web [19]. This causes low amount of recall with content mining.

## . V. WEB USAGE MINING

With the increasing demand of internet, more number of websites is being involved for getting required information and thus more usage of web-based data.

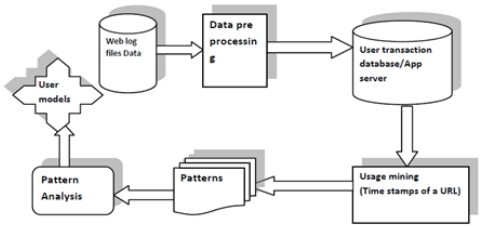
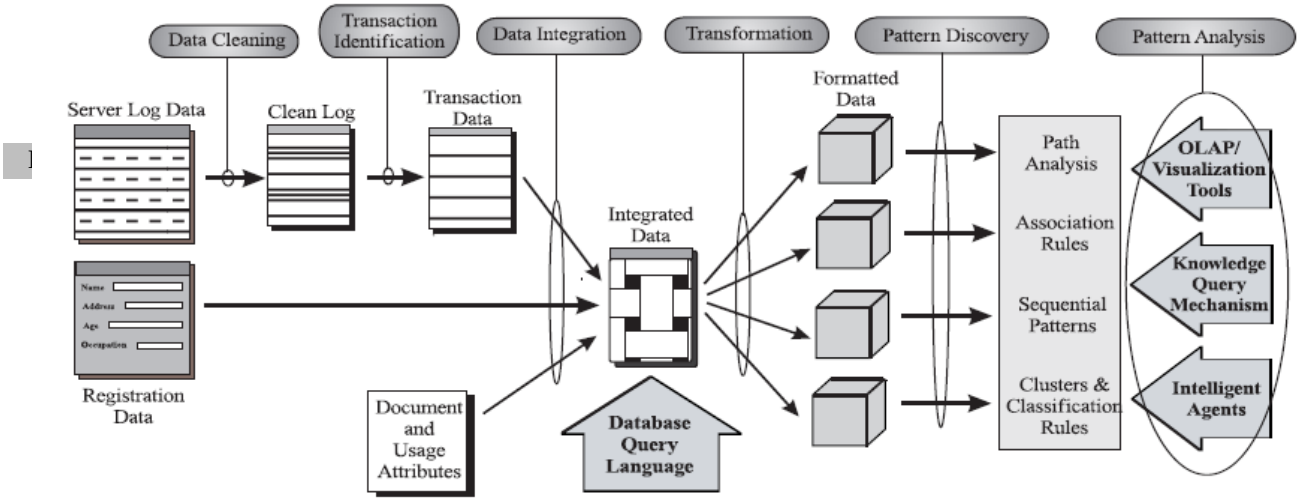



Figure 5. Web Usage Mining

The data that is stored in different format types in the web log file. This log file should be maintained as these data are in unsorted manner and it is done through preprocessing. Web usage mining focuses on discovering useful information. Web server automatically generates web log file whenever user accesses the resource like webpage of website.




Web usage mining is the process of withdrawing the useful knowledge from the server logs. It is the application of data mining techniques to discover interesting usage patterns from Web data in order to comprehend and better serve the requirements of the Web-based applications. Web usage data note down the identity of the user and their browsing behavior at a particular Web site. Usage data can be documented in the form of log files.

A Web log is a file in which the server takes the knowledge/data each time a user requests a site from a particular server. A log file can be placed in three different locations i.e. web servers, web proxy server, user's browser.

**Web Server Log files :** The log file that resides in the web server notes the activity of the client who accesses the web server for a web site through the browser.

**Web Proxy Server Log files :** The intermediate server (medium of interaction) exists between the client and Web server. Therefore, if the Web server gets a request of the client via the proxy server then the entries to the log file will be the information of the proxy server and not of the original user. These web proxy servers keep a separate log file for gathering the information of the user.

**Client/User Browsers Log files :** These log files can be made to reside in the client's browser window itself. A number of software's are there that can be downloaded by the user to their browser window. Even then, the log file is present in the client's browser window, only the Web server does the entries to the log file.

Web usage mining itself can be classified further depending on the kind of usage data considered:

**A. Web Server Data:** the Web server collects the user logs. Typical data includes IP address, page reference and access time.

**B. Application Server Data:** Commercial application servers have significant features to enable e-commerce applications to be built on top of them with little effort. A key feature is the ability to track various kinds of business events and log them in application server logs.

**C. Application Level Data:** New kinds of events can be defined in an application, and logging can be turned on for them thus generating histories of these specially defined events. It must be noted, however, that many end applications require a combination of one or more of the techniques applied in the categories above [2].

### 5.1 WEB USAGE PHASES

Web Usage Mining consists of four basic steps, Data Collection, Data Preprocessing, Pattern Discovery and Pattern Analysis.

*A.* **Data Collection**: This is the first step in which user's log data is collected from various sources. This includes only the relevant data that is to be collected. Data source can be gathered at the server-side, client-side, proxy servers, or obtain from an enterprise's database, which contains business data or consolidated Web data.

*B.* **5.1.2 Data Preprocessing:**

Some databases are insufficient, inconsistent and including noise. The data pre-treatment is to carry on a unification transformation to those databases and the database will become integrate and consistent, thus results the database, which may mine. In the data pre-treatment work, mainly include data cleaning, user identification, session identification and path completion. Data Preprocessing extracts text format data form log file and store clean data into database.

*C.* **Pattern Discoveries**

After the change of the data in the log file into a formatted data, the pattern discovery process is under gone. Pattern Discovery Tools apply techniques from data mining, machine learning, statistics and pattern recognition etc. In other words, Pattern Discovery finds pattern, Classify data by applying mining techniques.

*D.* **Pattern Analyses**

This is the final stage of Web Usage Analysis. Pattern analysis finds knowledge from the discovered pattern of the interesting patterns by eliminating the irrelevant patterns. Pattern Analysis involves the validation and interpretation of

the mined patterns. Validation can be used to remove the irrelevant patterns and to extract the interesting patterns from the output of the pattern discovery process. The output result is in mathematic form, which is not suitable for direct human interpretations. So, Visualization techniques are used to interpret the results. The most general ways of analyzing user access patterns are either by using a knowledge query mechanism on a database such as SQL or data cubes to perform OLAP operations. Visualization techniques, such as graphing patterns are used for an easier interpretation of the results.

### 5.2 RESEARCH ISSUES ON WEB USAGE MINING

Web usage mining involves number of data mining techniques. Due to which it faces several issues Problems are [10] Session identification from the user, Common Gateway Interface (CGI) data, Catching, Dynamic pages of the websites,
Robot detection and filtering, Transaction and merchant identification.

## 6. CONCLUSION

In this paper has discussed about the combination of the two fast-developing research areas Semantic Web and Web Mining. With the information overload, Web mining is a new and promising research issue to help users in gaining insight into overwhelming information on the Web. Also, research issues and challenges in web mining and provided detailed review about the basic concepts of web mining, web content mining, structure mining, usage mining, tools, and types. Classification, processes, the taxonomy and the function of Web mining. Several open research issues and drawbacks which are exists in the current techniques are also discussed. Workshops on Web mining have been already or will be held **to** discuss its principle, architecture and algorithm in several international conferences. In this paper, we present a preliminary discussion about Web mining, including the definition, the taxonomy and the function. This study and review would be helpful for researchers those who are currently pursuing their research in the domain of web mining.

### REFERENCES


[1] Mr. Dushyant B.Rathod, Dr.Samrat Khanna, "A Review on Emerging Trends of Web Mining and its Applications" ISSN: 2321-9939

[2] Usama Fayyad et al, "The KDD Process for Extracting Useful Knowledge from Volumes of Data", Communications of the ACM, Vol. 39, No. 11, Nov. 1996, pp. 27-34.

[3] P. Maes. Agents that reduce work and information overload. Communications of the ACM, 37(7):30–40,1994.

[4] H. N. Vu, T. A. Tran, I. S. Na, and S. H. Kim,"Automatic Extraction of Text Regions from Document Images by Multilevel Thresholding and kmeans Clustering," In Proceedings of IEEE/ACIS 14th International Conference on Computer and Information Science (ICIS), pp. 329–334, 2015.

[5] R. Lokeshkumar1, R. Sindhuja2, Dr. P. Sengottuvelan, "A Survey on Pre-processing of Web Log Filein Web Usage Mining to Improve the Quality of Data" International Journal of Emerging Technology and Advanced Engineering, ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 4, Issue 8, August 2014.

[6] AshirrK Kashyap, Iflah Naseem, Dheeraj Mandloi, "Web Mining an Approach to Evaluate the Web", International Journal of Scientific Research in Computer Science and Engineering, Vol.5, Issue.3, pp.79-85, 2017

[7] [Joy Shalom Sona, Prof. Asha Ambhaikar" A Reconciling Website System to Enhance Efficiency with Web Mining Techniques" International Journal Of Scientific & Engineering Research Volume 3, Issue 2, February-2012 1 ISSN 2229-5518.

[8] Sonia Sharma, Munishwar Rai, "Customer Behaviour Analysis using Web Usage Mining", International Journal of Scientific Research in Computer Science and Engineering, Vol.5, Issue.6, pp.47-50, 2017

[9] Ashish Kumar Garg, Mohammad Amir, Jarrar Ahmed, Man Singh, Sham Bansa," Implementation of a Search Engine" International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064.

[10] Jaideep Srivastava, "Web Mining: Accomplishments & Future Directions", University of Minnesota USA, srivasta@cs.umn.edu,

[11] Md. Zahid Hasan, Khawja Jakaria Ahmad Chisty and Nur-E-Zaman Ayshik, "Research Challenge in Web Data Mining", International Journal of Computer Science and Telecommunications Volume 3, Issue 7, July 2012

[12] D.Jayalatchumy, Dr. P.Thambidurai, "Web Mining Research Issues and Future Directions – A Survey", IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 14, Issue 3.

[13] Ananthi.J," A Survey Web Content Mining Methods and Applications for Information Extraction from Online Shopping Sites", International Journal of Computer Science and Information Technologies, Vol. 5 (3), 2014.

[14] http://www.slideshare.net/Tommy96/web-mining-tutorial R. Lokeshkumar1, R. Sindhuja2, Dr. P. Sengottuvelan, "A Survey on Pre-processing of Web Log File in Web Usage Mining to Improve the Quality of Data" International Journal of Emerging Technology and Advanced Engineering, ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 4, Issue 8, August 2014.

[15] Alberto Sillitti, Marco Scotto, Giancarlo Succi, Tullio Vernazza," News Miner: a Tool for Information Retrieval"

[17] Mamta M. Hegde, Prof. M.V.Phatak, "Developing an approach for hyperlink analysis with noise reduction using Web Structure Mining", International Journal of Advanced Research in Computer Engineering & Technology Volume 1, Issue 3, May2012.