INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X

(Volume2, Issue5)

Available online at: www.Ijariit.com

Web Usage Mining Tools & Techniques: A Survey

Satya Prakash Singh*
Dept.of Computer Science & Engineering
Madan Mohan Malaviya University of Technology
Gorakhpur (U.P.) INDIA
satyaprakash.singh271@gmail.com

Meenu
Dept.of Computer Science & Engineering
Madan Mohan Malaviya University of Technology
Gorakhpur (U.P.) INDIA
myself_meenu@yahoo.co.in

Abstract---The Quest for knowledge has led to new discoveries and invention. That leads to amelioration of various technologies. As years passed World Wide Web became overloaded with information and it became hard to retrieve data according to the need .Web mining came as a violence to provide solution of above problem. Web usage mining is category of web mining. Web usage mining mainly circulation with discovery and analyzing of usage patterns in order to serve the needs of web based applications. The web usage mining mainly consist of three stages: data preprocessing, pattern discovery and pattern analysis. This paper is focused with the study of different tools and techniques for web usage mining.

Keywords- Web Usage Mining, Data Preprocessing, Pattern Discovery, Pattern Analysis.

I. INTRODUCTION

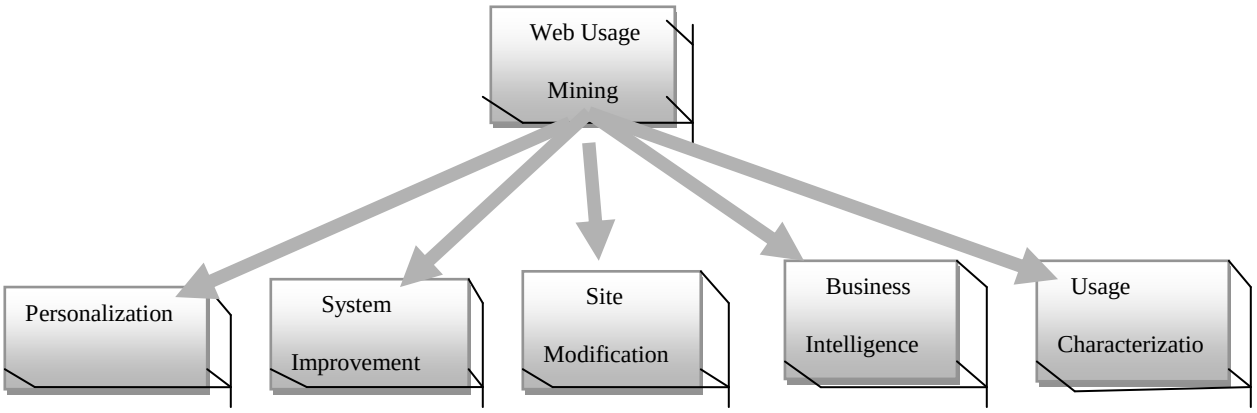
The advancement in the growth technology at a faster pace, world wild web has also grown exponentially. It has spread its arms along the entire world and being used in every field of day to day life. The World Wide Web is officially defined as a "wide- area hyper media information retrieval initiative aiming to give universal access to a large universe of documents". In simple term, the web is an internet-based computer network that allows users of one computer to access information stored on another through the world wide network called the internet.

The web mining technique acts as device to bring out of this trouble. It helps in automatic discovery and retrieval of information from the internet [10].Web Mining is divided into three categories: Web Content mining, Web Structure Mining and Web Usage Mining [14].Extracting useful information from the structured or unstructured contents of web document is described in Web Content Mining [4].In web structure mining, mining is done based on the structure like hyperlinks. In the case of web usage mining, mining is done on web logs which contain the navigational pattern of user.

Web usage mining focuses on techniques that could predict user behavior while the user interacts with the web. As mentioned before, the mined data in this category are the secondary data on the web as the result of interactions. These data could range very widely but generally we could classify then into the usage data that reside in the web clients' proxy servers and servers [12].

The structure of the paper is follows: section 2 presents the overview of web mining, web usage, its procedures and various techniques used. Section 3 deals with the literature survey and gives a brief of the recent researches done in the field of web usage mining. Section 4

enlightens the privacy issues related to web usage mining, Section 5 gives the description of web usage mining tools and their features, Section 6 and 7 describes the conclusion and future scope of the field and section 7 is about the references used.



FigureI. Web Usage Mining

II. WEB MINING

A. Overview

Web mining is the use of data mining techniques to automatically discover and extract information from web documents and services [16].This area of research is so huge today partly due to the interests of various research communities, the tremendous growth of information sources available on the web and recent interest in e-commerce. This phenomenon partly creates confusion when we ask what constitutes web mining and when comparing research in this area. Two different approaches were taken in initially defining Web mining. First was a 'process-centric view', which defined web mining as a sequence of tasks [14]. Second was a 'data-centric view', which defined web mining in terms of the types of web data that was being used in the mining process [4].

We decompose the web mining to following activity [2].

- Resource Finding: the task of retrieving indented web document.
- Information Selection and Pre-processing: automatically selecting and pre-processing specific information from retrieved web resources.
- Generalization: automatically discovers general patterns at individual web sites as across multiple sites.
- Analysis: validation and /or interpretation of the mined patterns.

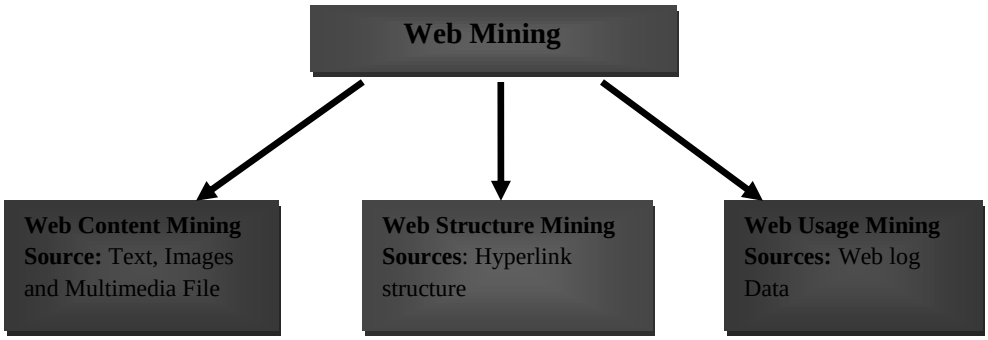
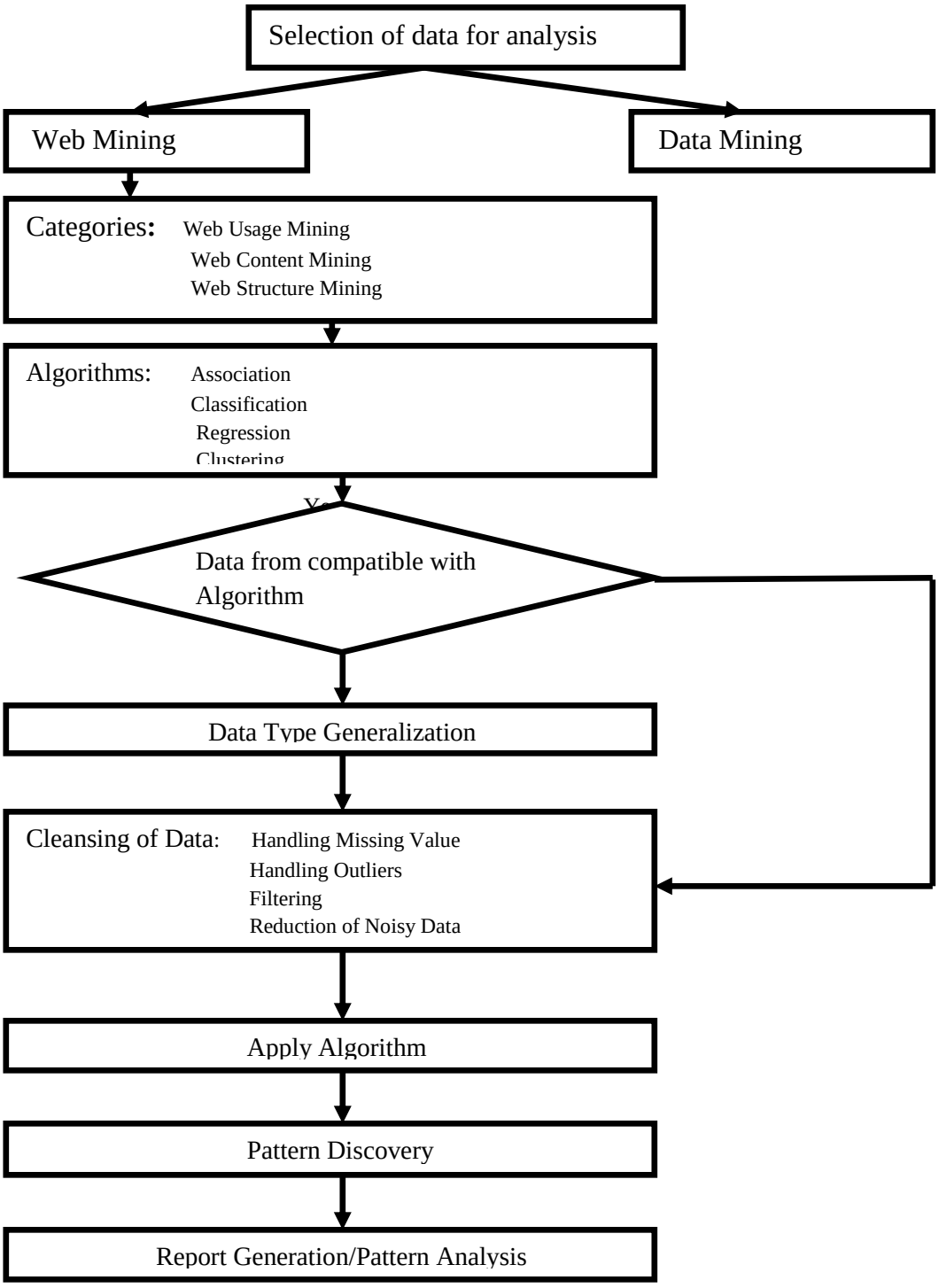


Fig II. The types and sources of Web mining

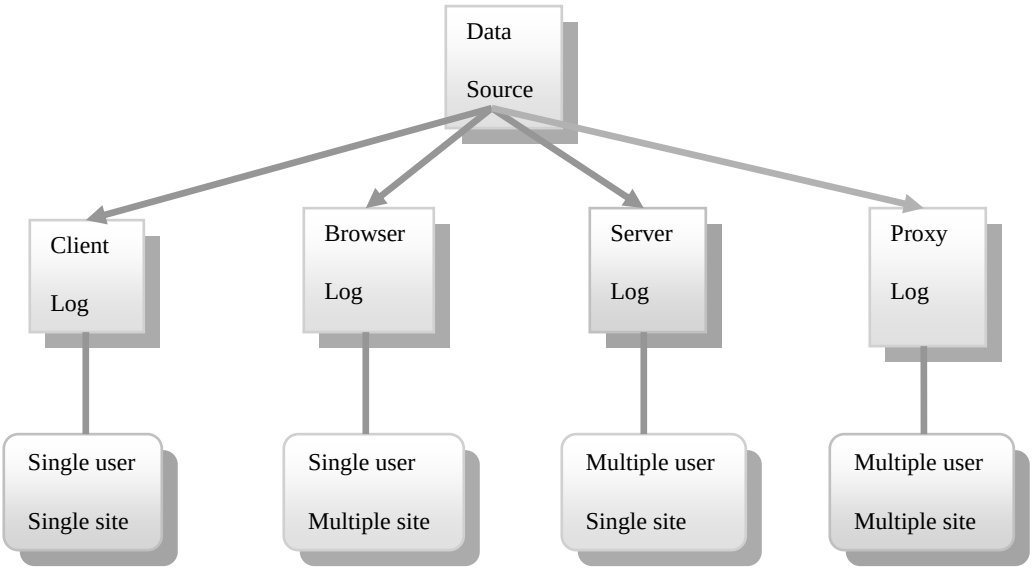
The Fig. Shows the types and sources of web mining. Web Content Mining is the process of extracting useful information from the contents of Web documents. Content data corresponds to the collection of facts a Web page was designed to convey to the users.



FigIII.Flow of Web mining

B. WEB USAGE MINING

Web usage mining is focuses on the distilling priceless intelligence from weblogs that can be proxy server logs, browser logs or merely log files. It center on click stream data, user profiles, registration data and numerous other data that contribute in a prediction of user purchasing or accessing behavior[3].Web usage mining is a technique of web mining which is based on the patterns we get from various weblogs and from these patterns we classified users.



FigIV. Various Data sources for Web Usage Mining

1. Client Level Collection

Client level collection is means of java scripts or java applets. This data show the behavior of a single user on single site. Client side data collection requires user participation for enabling java scripts or java applets.

2. Browser Level Collection

Browser level collection of the data collection is by modifying the browser. It shows the behavior of single user over multiple sites. The data collection capabilities are extend by modifying the source code of existing browser. Browser provides much more versatile data as they consider the behavior of single user on multiple sites [15].

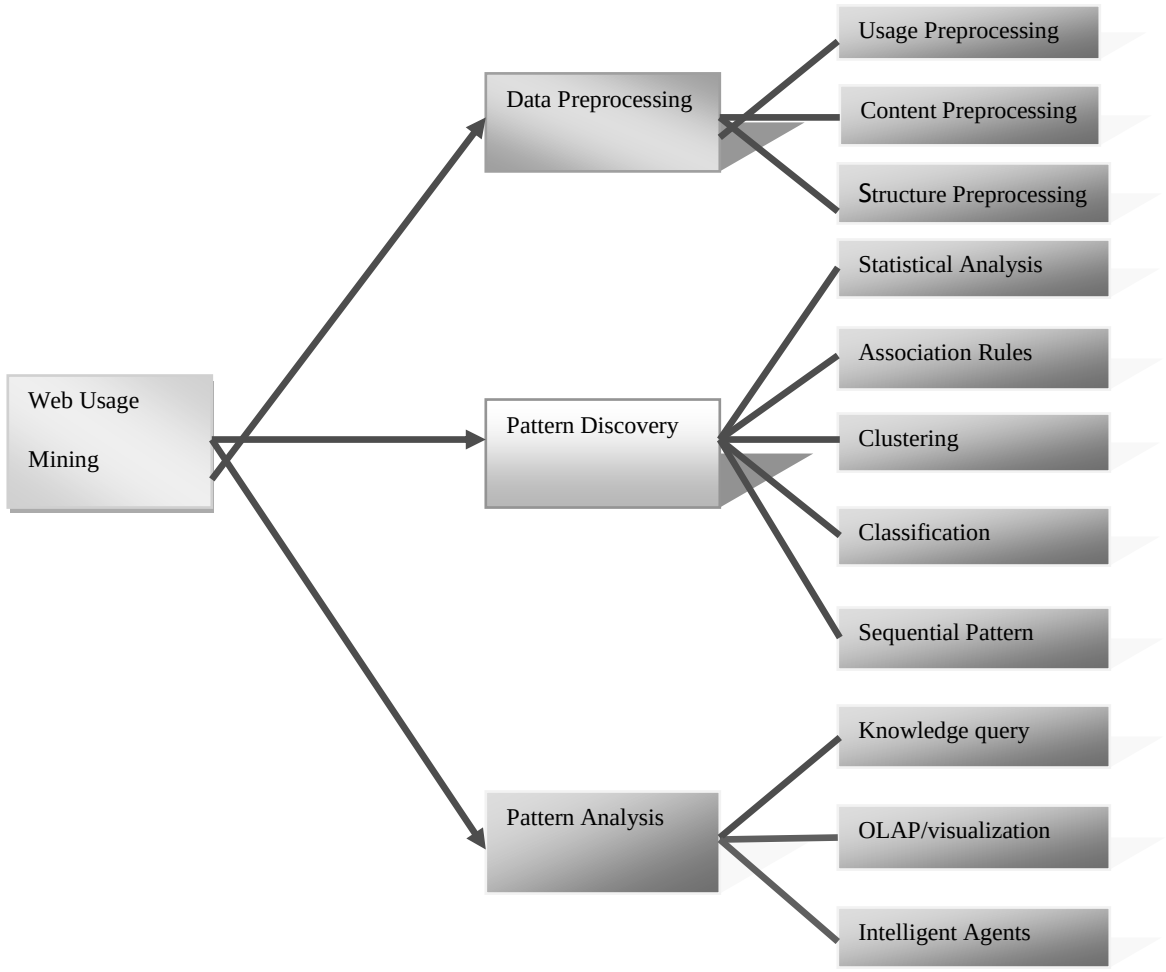
3. Server Level Collection

Server level collection behavior of multiple users over single site. Server log files can be stored in common log format or extended log format. Server logs are not able to store cached page views. Another technique used for usage data collection at server level is TCP/IP packet sniffing.

4. Proxy level collection

Proxy log servers are used by internet service provider to provide World Wide Web access to customers. These server stores the behavior of multiple user at multiple site. Proxy log functions like cache server and they are able to produce cached page views.

C. Web Usage Mining Procedure and Techniques



FigV.Web Usage Mining Procedures and Techniques

1. Data Preprocessing

Data preprocessing consists of converting the usage, content and structure information contained in the various available data sources into the data extract necessary for pattern discovery. Data preprocessing is categorized into three types: usage preprocessing, content preprocessing and structure preprocessing.

a. Usage Preprocessing

Usage preprocessing is arguably the most difficult task in the web usage mining process due to the incompleteness of the available data. Unless a client side tracking mechanism is used, only IP address, agent, and server click stream are available to identify users and server session.

Some of the typically encountered problems are:

- **Single IP address / Multiple Server Sessions: -** Internet service providers (ISPs) typically have a pool of proxy servers that users access the web through.
- **Multiple IP address / Single Server Session: -** Some ISPs or privacy tools randomly assign each request from a user to one of several IP addresses. In this case, a single server session can have multiple IP addresses.
- **Multiple IP address / Single User: -** A user that accesses the web from different machines will have a different IP address from session to session.
- **Multiple Agent / single user: -** Again a user that uses more than one browser, even on the same machine, will appear as multiple users.

b.Content Preprocessing

Content Preprocessing concerned with transforming unstructured and semi structured documents into the forms that are suitable for web usage mining. It is used for limiting the discovered pattern for web usage mining. Vector space model [5] is applied on page views in order to convert than into suitable format. The content of each page view to be preprocessed must be "assembled", either by an HTTP request from a crawler, or a combination of template, script, and database accesses.

c.Structure Preprocessing

Structure preprocessing of site is created by the hypertext links between page views and the frame and image tags that populate a particular page view. Several usage preprocessing steps cannot be completed without the site structure.

2. Pattern Discovery

Pattern discovery makes upon methods and algorithms developed from several fields such as statistics, data mining, machine learning and pattern recognition. Discovery of desired patterns and to extract understandable knowledge from them is a challenging task.

a. Statistical Analysis

Statistical analysis is the most common method to extract knowledge about visitors to a web site. The analyzing the session file, one can perform different kinds of descriptive statistical analyses (frequency,mean,median) on the variables such as page views, viewing time and length of a navigational path.

b.Association Rule

Association rule are an important class of regularities in data. Mining of association rules is a fundamental data mining task. It is perhaps the most important model invented and extensively studied by the database and data mining community. Its objective is to find all co-occurrence relationships, called association, among data items.

c.Clustering

The main purpose of clustering in web usage mining is to aggregate the similar sessions together [11], [17].Self organized maps, graph partitioning, and based technique, K-means with genetic algorithms, EM-CFuzzy means algorithm are the algorithms used for clustering the session. Types of clustering's.

- A clustering is a set of clusters.
- One important distinction is between hierarchical and partitioned sets of clusters.

d.Classification

The supervised predicting algorithm that exercise if than conjuction.It's a class-based approach in which lately encountered pattern is labeled according to the predefined class. Decision Tree induction is a trendy method of classification [7].

e.Sequential Pattern

Sequential pattern discovery attempts to find inter-session patterns such that the presence of a set of items is followed by another item in a time-ordered set of session or episodes. Given a set of objects, with each object associated with its own timeline of event, find rules that predict strong dependencies among different events.

$$(AB)(C) \Rightarrow (DE)$$

3. Pattern Analysis

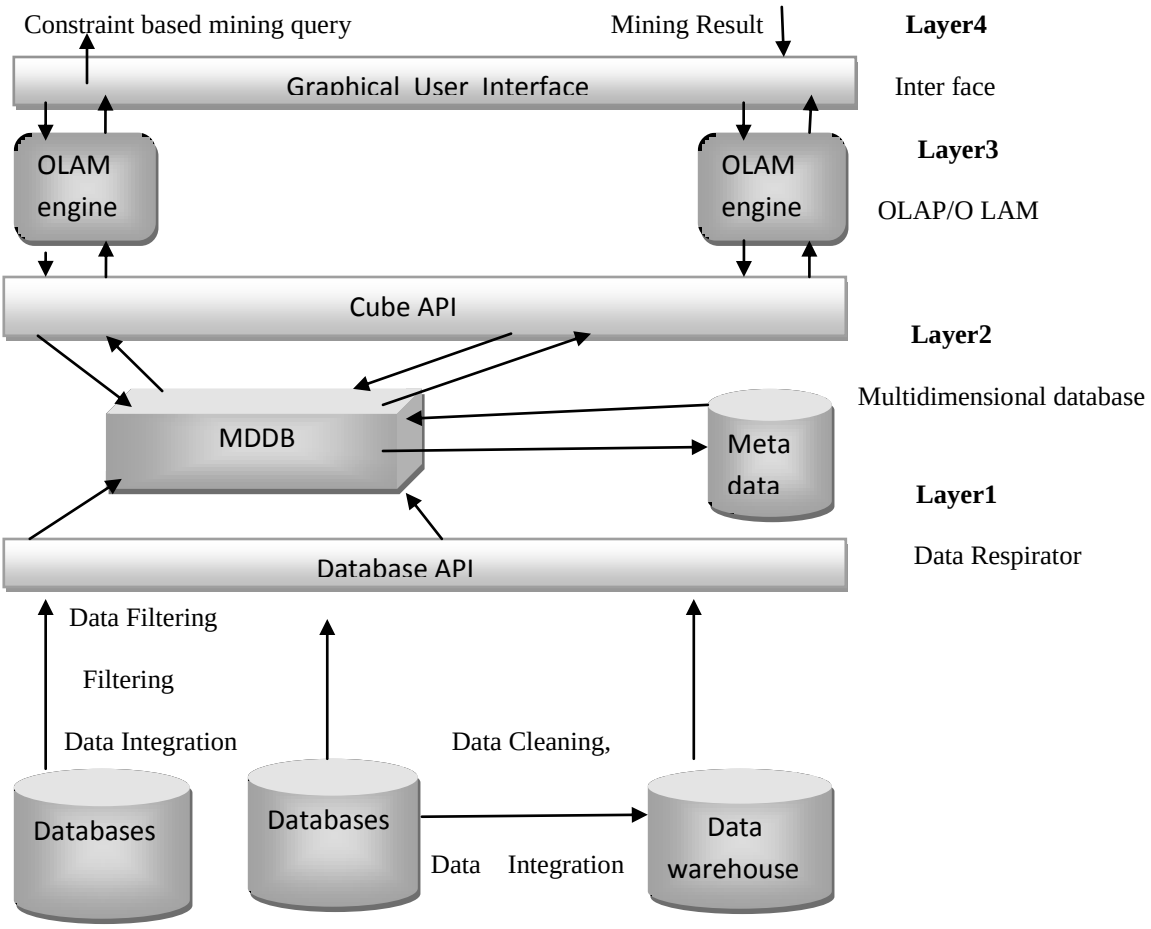
Pattern analysis is the last step which leads to achievement of web usage mining. It involves filtering out the create of the pattern discovery phase by expulsion monotonous patterns. Pattern analysis consist of knowledge discovery of exciting trends from the entire the available trends, the back end of this development is structured Query language [6].Pattern can be analyzed by using following techniques described below:

a. Knowledge Query Mechanism

Knowledge query mechanism is most commonly used language for structured query language (SQL) .This language is applied in order to extract the useful pattern from discovery patterns.

b.OLAP/Visualization

The OLAP tools in which discovered facts are placed on the cubes for performing various OLAP operations such as roll up and drill down and interesting facts are retrieved. Online Analytical Mining integrates with Online Analytical Processing with data mining and mining knowledge in multidimensional database. Here is the diagram that shows the integration of both OLAP and OLAM.



FigVI.Integration of both OLAP and OLAM

c.Intelligent Agent

Intelligent agents are also devised that helps in examining the patterns in web usage mining. These agents perform the work of analyzing the discovered patterns

III. RELATED WORK

Web usage mining system is divided into two queues of tracking and analysis in order to find out the user access pattern. The resultant of web usage mining which is knowledge about user access pattern is useful in various applications. We personalization is introduced as a solution [23].A complete framework for web usage mining and FM model is introduced in order to analyze the access pattern of user. The framework has been capable in doing implicit tracking which handle sparsely problem in web personalization and capturing short term change in user behavior. Web usage mining is being used in various spheres. In [13] web usage mining is used for improving the scalability and answer time of search engines.

Relationship between web usage mining and web structure mining was discussed in [9].Integration of both the techniques provides the advantage of faster web access to user, saving server memory space and better bandwidth utilization. A review has been done on recent developments in web usage mining research and is discussed in [8]. The process, techniques and applications of web usage mining is explained.

Web log record are used to discover the user entrance patterns through the help of users behavior and session [3].Discovery of browser pattern for getting data object into cache before an external request is made for a performance in terms of searching and web assessing[6].

For generating item recommendations by using lexical patterns, an action based rational recommendation technique is proposed [1]. And different pattern discovery techniques used by different researchers.

Web-page recommendation: - Recommender system allows to learn user preferences and to make recommendation. They can be employed to recommend products (e.g., books, movies, music, etc) and web content (e.g., news, photo, etc). Recommender system can be :(i) content based, the system recommends items similar to the ones the user preferred in the past, (ii) collaborative-filtering based, the system recommends items that people with similar tastes liked in the past, and (iii) hybrid, the system combines content and collaborative-filtering based method.

IV. PRIVACY CONCERNS

The increase in the growth of e-commerce on the web, the need for maintaining the privacy of the user while using any site also increases. Global and self regulatory nature of web raises the need for improving the privacy bar of user. To deal with this issue W3C introduced platform for privacy Preference (P3P)[9].A protocol is provided by P3P which allows site administrator to publish privacy policies , and when user visit site he has to agree with these policies in order to use site.

V. WEB USAGE MINING TOOLS

Web usage mining are different types of tools used in all the three stages of web usage mining are described.

TABLE 1
TOOLS USED IN VARIOUS STAGES OF WEB USAGE MINING

TOOLS	**FEATURES**
Data Preprocessing Tools	
Data Preparator	Performs cleaning, extraction and transformation of data before pattern discovery.
.Sumatra TT	Platform independent data transformation tool. Based on Sumatra script and support Rapid application Development.

Lisp Miner	Performs data preprocessing by analyzing the click stream and data collected.
Speed Tracer	Mines web server logs and reconstruct the user navigational path for session identification.
Pattern Discovery Tools	
SEWEBAR-CMS	Provides interaction between data analyst and domain expert to perform discovery of patterns. Helps in selection of rules among various rules in association rule mining [34].
i-iner	Discovery data cluster by using fuzzy clustering algorithm and fuzzy inference system for pattern discovery and analysis [33].
Argunaut	Develop the patterns of useful data by using sequence of various rules.
MiDas(Mining Internet Data for Associative Sequences)	Discover marketing based navigational pattern from log files. It applies more features to traditional sequential method.
Pattern Analysis Tools	
Webalizer	GNU GPL license based and produces web pages after analyzing patterns.
Naviz	Visualization tool that combines 2-D graph of visitor access and grouping of related pages. It describes the pattern of user navigation on the web.
WebViz	Analyze the patterns and provides them in the form of graphical patterns.
Web Miner	Mines the useful patterns and provides the user specific information.
Stradyn	Enhances WUM and provides visualization of patterns

VI. CONCLUSION

This paper has attempted to provide an up-to-date survey of rapidly growing area of web usage mining. With the growth of the web based tools and techniques, web usage mining is used in various area such as e-business, e-CRM and digital libraries and so on. Content and structure preprocessing allows raw data to be preprocessed along these dimensions also. Patterns are discovered by making various techniques like statistical analysis, association rule, clustering, classification, sequential pattern. They include of knowledge query mechanism and intelligent agent improvement the efficiency of pattern analysis.

REFERENCES

[1] P. Lopes and B. Roy, "Dynamic Recommendation System Using Web Usage Mining for E-commerce Users," Procedia Comput. Sci., vol. 45, pp. 60–69, 2015.

[2] Mele, Ida. "Web usage mining for enhancing search-result delivery and helping users to find interesting web content." In *Proceedings of the sixth ACM international conference on Web search and data mining*, pp. 765-770. ACM, 2013.

[3] S. G. Langhnoja, M. P. Bardot, and D. B. Mehta, "Web Usage Mining Using Association Rule Mining on Clustered Data for Pattern," Int. J. Data Min. Tech. Appl., vol. 2, no. 1, pp. 141–150, 2013.

[4] Sharma, Kavita, Gulshan Shrivastava, and Vikas Kumar. "Web mining: Today and tomorrow." In Electronics Computer Technology (ICECT), 2011 3rd International Conference on, vol. 1, pp. 399-403. IEEE, 2011.

[5] Han, Jiawei, Jian Pei, and Micheline Kamber. Data mining: concepts and techniques. Elsevier, 2011.

[6] S. R. Aghabozorgi and eh Y. Wah, "Using incremental fuzzy clustering to web usage mining," in IEEE International Conference on Soft Computing and Pattern Recognition, 2009, pp. 653–658.

[7] X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J.McLachlan, B. L. Angus Ng, P. S. Yu, Z.-H. Zhou, M. Steinbach, D. J.Hand, and D. Steinberg, "Top 10 algorithms in data mining," Knowl.Inf. Syst., vol. 14, no. 1, pp. 1–37, 2008.

[8] G.R.B., S.Totad, P.PVGD."Amalgamation of Web Usage Mining and Web Structure Mining.", International Journal of Recent trends in Engineering Vol.1 No2 May 2009.

[9] R. Ivancsy, I. Vajk.,"Frequent Pattern Mining in Web Log Data", Acta Poly-technical Hungarica vol3 No1 2006.

[10] Eirinaki, Magdalini, and Michalis Vazirgiannis. "Web mining for web personalization." ACM Transactions on Internet Technology (TOIT) 3, no. 1 (2003): 1-27.

[11] J. Heer and Ed H. Chi.,"Mining the Structure of User Activity Using Cluster stability", In Proceedings of the Workshop on Web Analysis, Second SIAM Conference on Data Mining ACM Press, 2002.

[12] Srivastava, J., R. Cooley, M. Deshpande Mukund, and P. N. Tan. "Web usage mining: discovery and application of usage patterns from web data." Proceedings of SIGKDD explorations 1, no. 2 (2002).

[13] J. Kerkhofs, K. Vanhoof and D. Pannemans,"Web Usage Mining on Proxy servers: A Case Study", Limburg University Center, July 30, 2001.

[14] Kosala, Raymond, and Hendrik Blockeel. "Web mining research: A survey." ACM Sigkdd Explorations Newsletter 2, no. 1 (2000): 1-15.

[15] Srivastava, Jaideep, Robert Cooley, Mukund Deshpande, and Pang-Ning Tan. "Web usage mining: Discovery and applications of usage patterns from web data." Acm Sigkdd Explorations Newsletter 1, no. 2 (2000): 12-23.

[16] O. Etzioni. The World Wide Web: Quagmire or gold mine. Communications of the ACM, 39(11):65–68, 1996.

[17] U. Fayyad G. Piatetsky-Shapiro and P. Smyth.,"From Data Mining to Knowledge Discovery: An Overview", In Proc. ACM KDD 1994.

