INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X

Impact factor: 4.295

(Volume3, Issue1)

Available online at: www.ijariit.com

A Survey on Pattern Discovery of Web Usage Mining

Manoj Kumar
Computer Science and Engineering
Madan Mohan Malaviya University of Technology Gorakhpur India.
manojkumaritmgida@gmail.com

Mrs. Meenu
Computer Science and Engineering
Madan Mohan Malaviya University of Technology Gorakhpur India.
myself_meenu@yahoo.co.in

Abstract --In the recent years with the development of Internet technology the growth of World Wide Web exceeded all expectations. A lot of information is available in different formats and retrieving content has become a very difficult task. One possible approach to solve problem is Web Usage Mining (WUM).Web mining is the application of data mining on web data and web usage mining is an important component of web mining. The goal of web usage mining is to understand the behavior of web site users through the process of data mining of web data and Web usage mining is to understand the behavior of web site users through the process of data mining of web Access data. knowledge obtained from web usage mining can be used to enhance web design, introduce personalization service and facilitate more effective browsing the important an application of web mining extracting the hidden knowledge in the log files of a web server recognizing various interests of web users, discovering customer behavior while at the site are normally referred as the application of web usage mining. In this paper we provide an updated focused survey on different pattern discovery techniques of web usage mining.

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

1. INTRODUCTION

WWW is a very popular and interactive medium for propagating information today. Due to the vast, varied and dynamic nature of web it raises the scalability, multimedia data and temporal issues respectively. The development of the web has been rise to large quantity of data that is freely available for user accessed by different users effectively and efficiently. That is why; the number of researchers in the field of application of Data mining techniques on the web is increasing .The web mining is the set of techniques of data mining applied to extract useful knowledge and implicit information from web data . As more organizations rely on the internet to conduct daily business, he study of web mining techniques to discover useful knowledge has become increasingly important However, with the magnitude and diversity of available information from the internet, it is not insignificant to locate the relevant information to satisfy the requirements of the people with different background. To assist Web surfers in browsing the Internet more efficiently, one of the topics that have attracted much attention is modeling the web user's browsing patterns and making recommendations.
Web mining enables one to discover web pages, text documents, multimedia files, images and other types of resources from web.

2. Web Usage Mining(WUM)

WUM is that area of Web mining which deals with the application of data mining techniques to reveal interesting knowledge from the WUD. WUM is three phase processes [15] that include data collection and preprocessing, pattern analysis of web data. Web usage mining concentrates on the techniques that could predict the navigational pattern of the user while the user interacts with the web. It is mainly divided into two categories, they are general access pattern tracking and customized usage tracking. In general access pattern tracking information is discovered by using the history of web page visited by user while in customized usage tracking mining is targeted on specific user. Mainly there are four types of data sources present in which usage data is recorded at different levels they are: client level collection browser level collection, server level collection and proxy level collection as shown in fig 1.

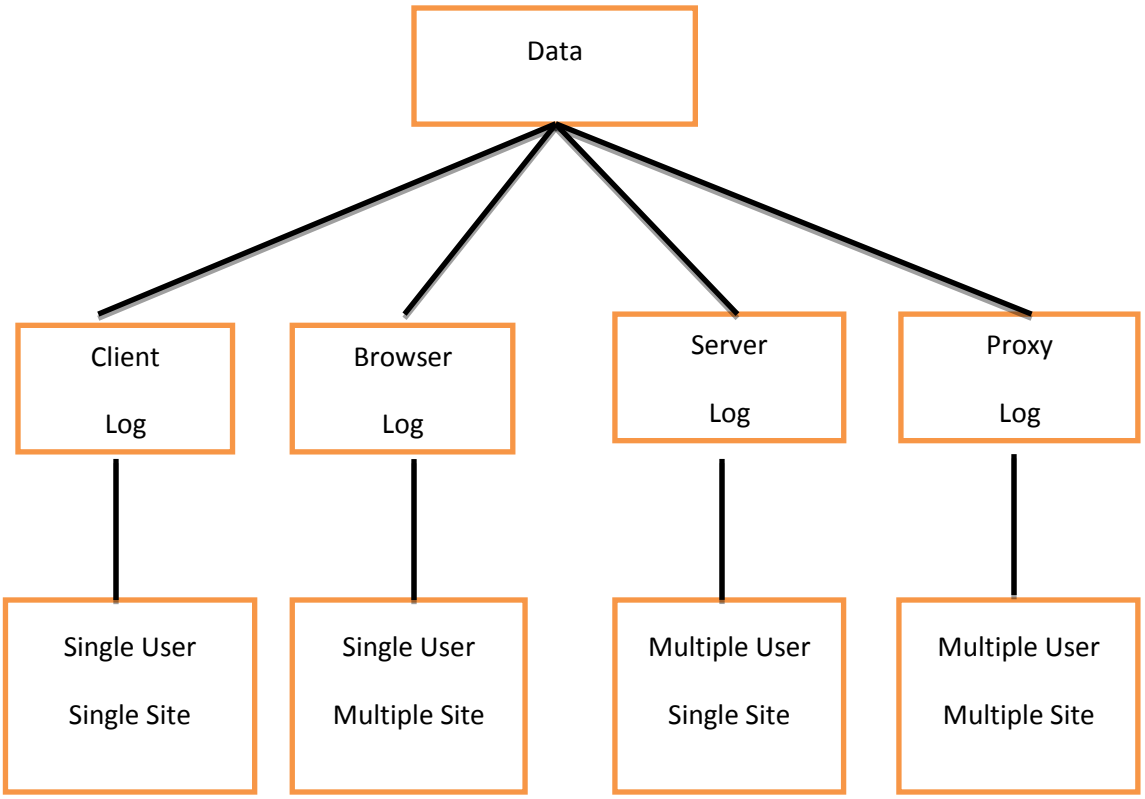

Fig. 1: Data Sources

2.1 Client level collection:

At this level data is gathered together by means of java scripts or java applets. This data shows the behavior of a single user on single site. Client side data collection requires user participation for enabling java scripts or java applets. The advantage of data collection at client side is that it can capture all clicks including pressing of back or reload button [1].

2.2 Browser level collection:

Second method of data collection is by modifying the browser. It shows the behavior of single user over multiple sites. The data collection capabilities are enhanced by modifying the source code of existing browser. They provide much more versatile data as they consider the behavior of single user on multiple sites [1].

2.3 Server level collection:

Web server log stores the behavior of multiple users over single site. These 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 sniffers works by monitoring the net-work traffic and retrieve usage data directly.

2.4 Proxy level collection:

Proxy 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. These server functions like cache server and they are able to produce cached page views by predicting the usage pattern of the visitor Web usage mining improves the quality of e-commerce services, personalizes the web [15]or enhances the performance of web structure and web server.

3. Web usage mining procedure and techniques

Three main steps: Data preprocessing, pattern discovery, and pattern analysis. This section presents an overview of each step and techniques used in them as shown in fig.2.

3.1 Usage preprocessing

This is considered as most difficult task of web usage mining because of presence of incomplete and inconsistent data in server log. Only IP address, agent and server side click stream are available to identify users and server sessions which faces many problems like single IP address/multiple server sessions multiple IP address /single server session, multiple IP address/single user and multiple agent/single user. Usage preprocessing also encountered the problem of inferring cached page references.

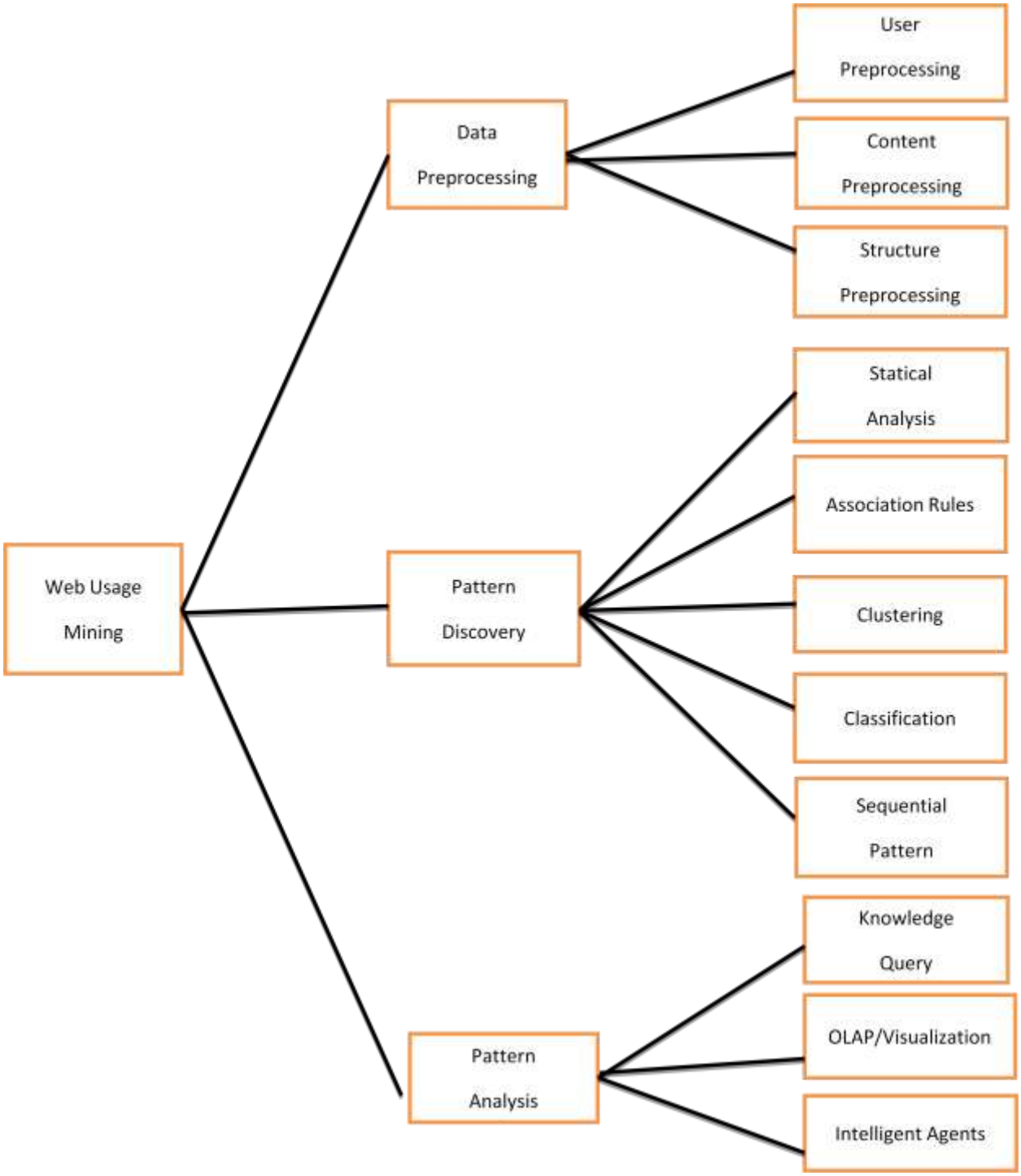

Fig.2-Web usage mining procedure and techniques

3.2Content 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 limited the dis-covered pattern for web usage mining.Vector space model[15] is applied on page views in order to convert them into suitable format. After proper formatting content mining algorithm are applied. Preprocessing of the content of each page view is performed either by HTML request or from a combination of template, script and database access [1].

3.3 Structure preprocessing

Structure of page views includes the hyperlinks. Structure preprocessing is performed similarly as content preprocessing is performed. In case of dynamic content a different site structure is developed for each session.

4. Pattern Discovery

Pattern discovery focuses on the uncover pattern from the abstractions produced as a result of preprocessing phase. It focuses on applying various methods and techniques developed from several fields such as data mining, machine learning, statistics and pattern recognition. Discovery of desired pattern and to extract understandable knowledge from them is a challenging task. This section explains some of algorithms suitable for pattern discovery

4.1Statistical Analysis

Statistical analysis focuses on analyzing the data which is aggregated by predetermined units such as days, sessions, visitors in order to gain knowledge about user behavior. Mainly three type of statistical analysis(frequency, mean, median) is performed on sessions and result of analysis shows frequently accessed pages, average view time or length of the page. These results become helpful in improving system performance or enhancing the security [1].

4.2Association Rules

Association rules are used for finding the correlations among web pages that frequently appear together in a user browsing session [13] is the most popular algorithm that expresses the frequent co-occurrence of web pages together. These rules help in providing the recommendations to user visit the sites and web designers to restructure web sites. These rules show that if page M is accessed by user them it is very likely that is visits page N also in same session. Algorithms used for association rules include maximal forward references, Markov Chains, FP growth and prefix span.

4.3 Clustering

The main purpose of clustering in web usage mining is to aggregate the similar session together [11],[12].Self organized maps, graph partitioning, ant based technique, K-means with genetic algorithms, EM-C Fuzzy means algorithms are the algorithms used for clustering the sessions. Mainly clustering is two types usage clustering and page clustering. Usage clusters discover user that have some browsing pattern where page cluster gather the content related page together. This is pre-formed by transforming the session into vectors of N elements where N may be no of pages or page views. Clustering helps in personalizing the web site as it identifies the users with similar behavior.

4.5 Classification

Classification is supervised way of learning which maps the data items to one of the many predefined classes. Various supervised learning algorithms used for doing classification are decision tree, naïve Bayesian classifiers, k-nearest neighbor classifiers and support vector machine. Classification mainly performs the automatic categorization of documents. In web usage mining, application of classification algorithms on server logs may lead to detection of interesting pattern such as 40% of users who visits news site are in the age group of 30-35 years.

4.6 Sequential pattern

In web usage mining sequential pattern is used to discover the sessions that are found in a sequence. They include the sequence of items that frequently occur in a particular order. MIDAS (Mining Internet Data for Association Sequences) algorithm is most commonly used for finding sequential pattern which provide marketing intelligent behavior for e-commerce scenario.

5. Pattern Analysis

The last step of web usage mining process is pattern analysis. This phase separates the interesting and uninteresting pattern from the overall pattern discovered during pattern phase. Result of pattern analysis these is used in various application such as system performance improvement, site modification, personalization, e-commerce etc. pattern can be analyzed following techniques described below.

5.1 Knowledge Query Mechanism

Structure Query Language(SQL) is most commonly used language for knowledge query mechanism. This language is applied in order to extract the useful pattern from discovered patterns.

5.2 OLAP/Visualization tools

Pattern are also analyzed by using OPAL tools in which discover facts are placed on to data cubes for performing variousOLAP operation such as roll up and interesting facts are retrieved. OLAP provides an integrated framework for analysis which allows changes in aggregate levels. Output of OLAP queries acts as an input to data mining or data visualization tools. Graphing patterns or assigning colors to different values are used as visualization technique for some purpose.

6. Literature Work

Web log records are used to discover the user entrance patterns through the help of users behavior and session [5].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 [2].user's activities while browsing the internet show their attention toward various content available online and also with E-commerce sites. Organization tends to provide better access to their user by capturing their purchasing interest and steering details. Path visited, access time and frequently browsed content are their interest. Combining these pillars to form a clustering algorithm which will group users with matching behavior. These set of result can be help various business plan to build an efficient E-commerce click flow is recorded for optimal outcome. Form the post experience it is known that user's active while browsing through internet show their interest to words various content available online and also with e-commerce sites. Organization lands to provides better access to their user by capturing their purchasing interest and navigation details .further storing information like path visited and time spend on each web page by the user [3] large number of queries floating to web server from user end lead to a very low response time so to enhance the resulting quality of optimizing searching engine frequent output of the query is stored in cache[4] customer contentment can be ornamental by providing location based web services (LBWS) by relating location-based information (LBI). This information is used as for providing fruitful
Executable web services provided through simple object access protocol. Relationship should be catch between LBI and LBWS for location based approach [5] the importance of catching constructive information and knowledge discovery from logs has been quite apparent. Web usage mining cover the data mining loop web mining through cloud computing will gain brilliant attention in future term as cloud mining [3] web mining originated from data mining offline analysis of behavior formulate data mining whereas online referred as web mining. Three segments preprocessing, pattern discovery and pattern analysis leads to web mining. Preprocessing remove anomalies by normalizing the web logs. Aprori and K-MEAN algorithm is used to computer frequent web pages access and unique user [6]. The number of visits on URL by a number of users can be formulated by using support and confidence parameter. To classify the related web pages visited collectively or relation between the chains of visited web pages.[7] for the betterment and fame of company knowledge of useful pattern is essential which is seized form logs leading to an analysis of website structure the value of larger forward references is obtained from maximal forward references. [8] due to not availably of logs publically it is difficult for researchers to gather knowledgeable patterns. Google chrome extension provides free access to private logs leading to the computation of user behavior by applying improved apriori algorithm [9]. Customer satisfaction is the main motto of web mining which can be accomplished by combing Association rule meaning and clustering algorithm i.e. DBSCAN algorithm. Clustering techniques are applied on outcome of association rule mining.[5] ample amount of web services are available due to which finding a suitable web service a method is proposed in which many web services that are inter related can be combined to gather which can be utilized by the user. the method uses a model named as semantic kernel model which discovers services that are related and then exhibits as a graph's node, and then an algorithm is used to find the best arrangement named as all pair shortest path. further empirical evaluation is used to check the accuracy of the method [10] mapping users on the basis of country, site entry to the websites and seasonal months can be helpful for grooming of websites by customer relevant needs location can be seized form IP address and mapping this to domain name server. month related seasonal details will be used to boost the pattern discovered [11] this research paper deals with the issues in making effective web page recommender using semantic enhancement. We further classified web page recommendation in two categories: sequence learning model (it involves several algorithms like tree-based algorithms, particularly pre-order linked WAP[-Tree mining (PL WAP-mine for short)semantic-enhanced approaches(it uses concept of domain ontology for classify pages and searching relevant pages).

Domain ontology deals with relationships amongst terms and represents domain knowledge for a specific domain.[12]

Estimated usage pattern is obtained by their user interactive path model on the basis of user behavior. Actual usage pattern is mined from logs of web servers which is regularly documented for every active website by firstly processing data of log to recognize users session, identity transactions then determining the pattern using algorithm of usage mining[13]. Combining different Markov model for low state complexity, accurate prediction and to filter surplus categories relevance factor is adopted to predict frequent strokes performed by a user which can be deliberate from the similarity of user attitude between web categories. In this the target is last frequent action performed depending upon the Markov model we decide the number of last actions to be considered to predict the outcome [14]Useful patterns and knowledge can be extracted by enhancing the traditional web usage mining techniques to be used focus of this paper is on providing real time dynamic recommendation to the website

Visitors whether they are registered are not. For generating item recommendations by using lexical patterns, an action based rational recommendation technique is proposed [15].table 1 show different pattern discovery techniques used by different researchers.

Table1. Different criteria for pattern discovery

Sr. No.	References	Year	Focus on	Parameters	Technique	Examine	Experimental Result
1	Shang-Pin et Al [5]	2012	Relaning LBI and LBWS	LBI LBWS	Association Rule Sequential rule	Service usage pattern and weblogs	Location based relevant and satisfied information
2	Chen et al [3]	2013	Interest based click stream Data records	Visiting Path Browsing FrequencyAccess time	Leader Clustering Algorithm	Website topology Diversified Commodity categories	Efficient decision making for E-commerce site Customer satisfaction
3	Meleet al [4]	2013	Cache of frequent results	Relative Position Timestamp	Greedy approach Adopter Graph	Logs to find Early adopters	Better search engine performance Recommendation of relevant web pages
4	MurliManohar et al [8]	2014	Website architecture or connectivity of webpages	MFR Node-path table	MF SPRA	The weblogs for user traversal patterns	For better and interactive user interface
5	Bhargava et al [11]	2014	Country time, access, seasons site entry	IP Address Timestamp	Classification Algorithm	Discovered patterns from logs	Increase overall profit by fulfilling user needs
6	Pardeshi et al [9]	2014	Employing Google chrome extension	FIT FRT	Clustering Algorithm	Hit Graph Time Graph	Information gathering
7	Parekh et Al [6]	2015	Frequent web pages and unique user	Frequent Item set Controid	Apriori Algorithm K MEAN clustering	Preprocessing result from weblogs	User satisfied results

Conclusion

Data mining is the study of exploring patterns in huge volumes of raw data. The term Web mining has been used to refer to techniques that help us to find content of web and retrieve the user's interest and needs. This paper focuses on the comparison of mining algorithms. Web Usage Mining and various criteria are considered for pattern discovery. As huge data is adding in repository every second so there is need of quality information to satisfy upcoming needs of user. In future Web Service provisioning and web service discovery should be made by analyzing user interest.

References

[1] K. Sharma, G.Shrivastava, and V. Kumar, "Web mining: Today and tomorrow," in IEEE 3rd International Conference of Electronics Computer Technology (ICECT), 2011, pp. 399–403.

[2] 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.

[3] L. Chen and Q. Su, "Discovering user's interest at E-commerce siteusing clickstream data," in IEEE 10th International conference onService systems andservice management (ICSSSM), 2013, pp.124–129.

[4] I. Mele, "Web usage mining for enhancing search-result delivery and helping users to find interesting web content," in ACM 6th International conference on Web search and data mining, 2013, pp. 765–770.

[5] S.-P. Ma and D.-Y. Yan, "Location-Based Web Service Delivery: Data Mining-Based Approach," in IEEE International Symposium on Computer, Consumer and Control (IS3C), 2012, pp. 666–669.

[6] A. Parekh, A. Patel, S. Parmar, and V. Patel, "Web usage Mining: Frequent Pattern Generation using Association Rule Mining and Clustering," Int. J. Eng. Res. Technol., vol. 4, no. 4, pp. 1243–1246,2015

[7] R. Agrawal T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases," ACM SIGMOD Rec., vol. 22, no. 2, pp. 207–216, 1993.

[8] M. M. Sharma and A. Bala, "An approach for frequent access patternidentification in web usage mining," in IEEE International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2014, pp. 730–735.

[9] S. Pardeshi and P. Patil, "Data‖: An Overview," Int. J. Adv. Res. Comput. Commun. Eng., vol. 3, no. 1, pp. 5117–5121, 2014.

[10] R. Nayak and A. Bose, "A Data Mining Based Method for Discovery ofWeb Services and their Compositions," Real World Data Min. Appl., vol. 17, pp. 325342, 2015.

[11] A. Bhargav and M. Bhargav, "Pattern discovery and users classification through web usage mining," in IEEE International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT), 2014, pp. 632–636.

[12] T. T. S. Nguyen, H. Y. Lu, and J. Lu, "Web-page Recommendation based on Web Usage and Domain Knowledge," IEEE Trans. Knowl. Data Eng., vol. 26, no. 10, pp. 2574–2587, 2014.

[13] R. Geng and J. Tian, "Improving web navigation usability by comparing actual and anticipated usage," IEEE Trans. Human-Machine Syst., vol. 45, no. 1, pp. 84–94, 2015.

[14] V. M. Rao and V. V. Kumari, "An Efficient Hybrid Successive Markov Model for Predicting Web User Usage Behavior using Web Usage Mining," Int. J. Data Eng., 2010.

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

