A SURVEY: WEB USAGE MINING & RECOMMENDATION SYSTEM

Himanshi Kirar[1], Punit KumarJohari[2]

Abstract

From it is very beginning, the capability of finding relevant information from the Web has been quite apparent. Web mining (WM) – i.e. the type of data mining techniques to fetch useful information from Web content, structure, and usage is the collection of skills to fulfill this requirements. WM is the application of data mining strategies to fetch knowledge from Web data, where minimum one of structure (hyperlink) or usage (Web logs) data is used in the mining process (with or without other types of Web data). Interest in WM has grown quickly in its short existence, both in the research and practitioner communities.

Keywords: Web Usage Mining (WUM);KNN;Web Mining (WM)

I. INTRODUCTION

WUM is also known as web log mining is the application of data mining methods on large web log storehouses to discover useful knowledge about user's behavioral patterns and website usage statistics that can be used for several website design tasks. The main source of data for WUM consists of textual logs collected by numerous web servers all around the world. There are four stages in WUM.

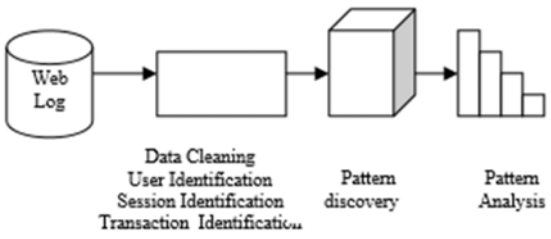

Fig.1 Phases of Web Usage Mining [1]

- Data Collection: Users log data is collected from various sources like server side, client side, and proxy servers and so on. [1]
- Preprocessing: A way performs a sequence of processing of web log file cover data cleaning, user identification, session identification, path completion and transaction identification. [1]
- Pattern discovery: In pattern discovery different types of data mining methods applied on the web data. Application of different data mining techniques/method to processed data as statistical analysis, association, clustering, pattern matching, classification and so on. [1]
- Pattern analysis: Once patterns were discovered from web logs, uninteresting and irrelevant rules are filtered out and relevant patterns uses in the form of web data. Analysis is done using knowledge query mechanisms like as SQL or data cubes to perform OLAP operations. All the four phases are delineated through the accompanying figure. [1,2]

The aim of this paper is to provide a review of WUM and a survey of preprocessing stage. Data Collection section list out various data sources, preprocessing section reviews the different works done in session identification, path completion process. Remaining sections briefs about pattern discovery, analysis and the different areas of applications where WUM is used [1].

II. TYPES OF WEB MINING

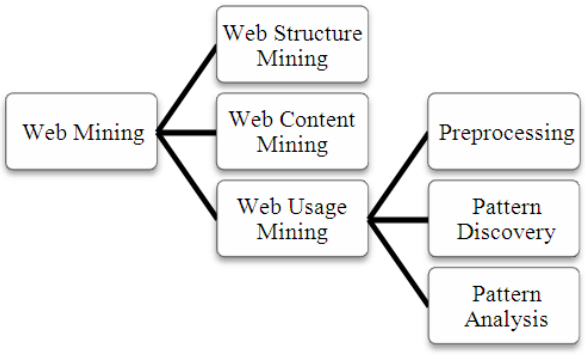

Fig.2 Categorization of Web Mining [13]

A. Web Content Mining

Web content mining is the extraction of useful abilities from the content materials, on web pages [6]. Much interesting information can be found by mining, web page content such as, item description, discussion posting etc. for several reasons.

Fig.3 Example of Web Content Mining [3]

B. Web Usage Mining

WUM can be defined as the discovery of consumer entry pattern from log file to find person behavior. WUM uses many data mining algorithms to perform its task. Preprocessing of web log is the key issues in usage mining. In addition to this WUM is further separated into three stages, namely data collection and data preprocessing, pattern mining and pattern analysis. [5, 2]

C. Web Structure Mining

Web structure mining refers to the extraction of valuable learning from hyperlink to WWW.Numerous helpful information can be found from the hyperlink structure of the web, for example, relationship between website pages, which pages are relatively more essential. This information further can be used by web master for intelligent decision-making. As stated via junior web structure mining specializes in the hyperlink constitution of the net and the different objects are linked by some means. Web structure mining could be further isolated into hyperlink and document structure. Hyperlink could be within a document or outside it. If it is within the document, then it is termed as Intra-document hyperlink if it is between two documents, then it is called Inter-document hyperlink. Moreover, all the techniques of WM described above cannot be applied directly on web data due to inappropriateness of data such as presence of noise, missing references etc. So it has to be passed through data preprocessing stage, which makes it suitable for mining and analysis task [1, 2, 6].

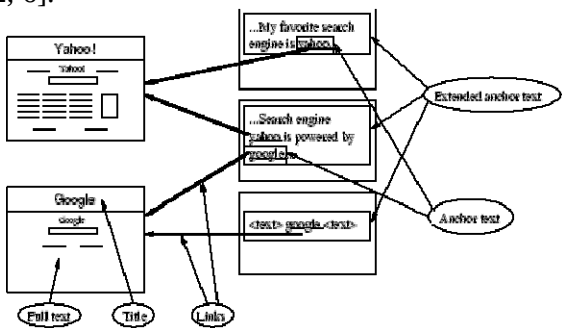

Fig.4 Example of Web Structure Mining[4]

III. CHALLENGES IN WEB MINING

The upper section emphasizes the detail that the WM system comes a long way and it is broadly utilized in dissimilar areas of research. A summarized seen in the previous area reveals the unfolded challenge which is listed below:

A. Unstructured Information

To manage the unstructured information in the web is one of the unsolved issues among WM system. The main cause for this unanswered problem is their weak operational techniques, means those systems and related tools, which have established to effectively changing over organized data into a data intelligence that systems are ineffective when we use to execute the same on unstructured information.

B. Semantic Interpretability

WM is a knowledge retrieval infrastructure, which exploits data mining technique to automatically extract wanted data in heterogeneous environment similar to the web. Currently human are able to understand the

unstructured query on web on the basis of their past experience and other side software agents are used by WM method to looking and recovering learning on the web. However, certain challenges are inhibiting in the implementation of software agents in WM infrastructure and a standout amongst the most noticeable challenges being the issue of semantic interpretability. Semantic interpretability rises the question, how to enable agents to acquire consistent knowledge from other agents, those agents are doing work in dissimilar domains and formulating use of different ontology [7].

IV. APPORACHES OF WEB USAGE MINING

A) Data Collection: Data collection is the first step of WUM, the data authenticity and integrality will directly affect the following works efficiently carrying on and the final recommendation on characteristic service's quality. Therefore, it must use reasonable and advanced technology to gather various data. Now days, towards WUM technology, the main data origin has three kinds: server data, client data and middle data [8].

B) Data Preprocessing: Some databases are inadequate, inconsistent and noisy. The data pretreatment is to carry on a unification transformation to those databases. The result is that the database will become integrated and consistent [8].

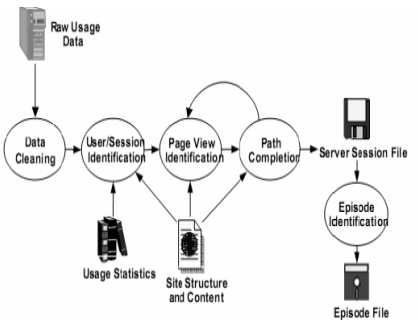

Fig.5 Preprocessing of web usage data

C) Knowledge Discovery: Use statistical method to carry on the analysis and mine the pretreated data. We may cover the user or the user community's interests, then construct interest model. At present the usually used machine learning methods mainly have clustering, classifying, the relation discovery and the order model discovery. Each method has its own particular brilliance and shortcomings, inadequacies, however the quite effective method mainly is classifying and clustering at the present [8].

D) Pattern Analysis: Challenges of pattern analysis are to purify uninteresting information and to visualize and decipher the interesting patterns to the user. First, delete the less interesting rules or models from the interested model storehouse; next use technology of OLAP and so on to carry on the comprehensive mining and analysis; Once more, let significant data or knowledge to be visible. Finally, provide the characteristic service to the electronic commerce website [8].

V. TECHNIQUE OF WEB MINING

Techniques: There are certain techniques which are used in web usage mining.

- Sequential-pattern-mining-based: This allows the new findings of the temporally ordered the Web access patterns [9].
- Association-rule-mining-based: This techniques is used to find the find relationship among different Web pages [9].
- Clustering-based: Here users from the groups which are having similar types of characteristics [9].
- Classification-based: Users are grouped into classes which are predefined and also based on some of their characteristics [9].

The KNN method is used online and also in Real-Time to increase web usage data mining technique to identify clients or visitors click stream data matching it to a particular user group and also to recommend a tailored browsing and searching option which meet the need of the specific user at a specified time. K-Nearest Neighbor (KNN) is depends on the rule that the occurrences within a dataset will generally exist in near proximity to other instances that have similar properties. As KNN does not make any assumptions on the primary data distribution and does not use the training data points to do any generalization, it is called as non- parametric lazy learning algorithm. Guo et al. [9] proposed a novel KNN type method for classification that is aimed at overcoming the downside of its dependency on the selection of a "good value" for k. Yu & Liu addressed the problem of determining which of the available input features

should be used in modeling via feature selection because it could improve the classification accuracy and scale down the required classification time [9].

This method, however, did not gain popularity until the 1960's with the availability of more computing power, since then it has become widely used in pattern recognition and classification. KNN learns by comparing a specific test tuple with a set of training tuples that are similar to it. It classifies based on the class of their closest neighbors and majority of neighbors, most often, more than one neighbor is taken into consideration, hence, the name K-NN, the ''K'' indicates the number of neighbors taken into account in determining the class. The KNN classification technique has been prepared to be used online and in real-time to recognize clients/visitors click stream data, matching it to a particular user group and recommend a customized browsing option that meet the need of the specific client at a specific time.

K-NN classifier for pattern recognition and classification in which a particular test tuple is compared with a set of training data that are similar to it. The K-NN algorithm is one of the simplest and easier methods for solving classification problems; it often yields competitive results and has important advantages over several other data mining methods.

(1) Providing a quicker and more precise recommendation to the web clients with desirable qualities as a result of straightforward application of similarity or distance for the purpose of classification.

(2) Our recommendation engine gathers all the active users' click stream data, match it to a specific user's group in order to generate a set of recommendation to the client at a faster rate.

VI. RECOMMENDATION SYSTEM

Web recommendation systems help the website visitors for easy navigation of web pages, quickly reaching their destination and to obtain relevant information. There are two types of approaches to develop recommendation system.

- Content based filtering method [10]
- Collaborative filtering method. [10]

In Content-based filtering technique filtering is done based on customer's interested items. In content-based filtering technique, the web pages are recommended for a user very quickly from ancient database. In that database different content of items are added that the user has used in the ancient times and/or user's personal information and preferences. The user's data files can be constructed by using responses to questions, item ratings, or the user's navigation information to infer the user's preferences and interests.

In collaborative filtering approach, web pages are recommended to a specific user when other similar kind of users also prefers those web pages.

VII. LITERATURE SURVEY

V. Anitha [2016] et al. gives an attention on WUM to predict the behavior of web users based on web server log files. Users using web pages, a frequent access path and frequent access pages, links are stored in web server log files. A Web log along with the individuality of the user captures their browsing behavior on a website and discussing about the behavior from analysis of different algorithms and diverse techniques. A presently internet is most imperative piece of human life. The internet is a growing day by day, so online users are also heightening. The interesting information for knowledge of extracting from such enormous data demands for new logic and the new method. Every user spends their most of the time on the internet and their behavior is different from one and another. The application of data mining techniques service to need of web-based applications. WUM is leading research area in WM concerned about the web user's behavior [11].

Meryem Boufim [2016] et al. presents that due to the exponential growth of internet using, the habits of internet users have changed. This generates huge amount of data. It becomes important to explore this mine of knowledge and take an advance on concurrent. In the context of digital marketing, the user's data is the enterprise assets to personalize the content of websites and establish contact and communication with customers through internet channels. The more the enterprise has data and knows about its non-real customers the more it has the advantage to take advance and be a leader. In this context inbound marking was recently born: to help marketers establishing customer-centric digital marketing strategy. On the other hand, WM is used to discover information through web data

and construct knowledge about online customers. The WM techniques seem to fit the best with the inbound marketing implementation. After an introduction to inbound marketing and web mining methods, this paper presents the application of WM methods and techniques to implement an inbound marketing strategy [12].

G. Neelima [2016] et al. present that It is the method to extract the user sessions from the given log files. Initially, each user is identified according to his/her IP address specified in the log file and corresponding user sessions are extracted. Two types of logs i.e., Server-part logs and customer-facet logs are in most cases used for web usage and usability analysis. Server-side logs can be consequently produced by web servers, with every passage comparing to a user request. Client side logs can seize correct, comprehensive usage data for usability analysis. Usability is defined because the delight, efficiency and effectiveness with which special customers can entire particular duties in a specified atmosphere. This procedure includes 3 phases, namely Data cleaning, User identification, and Session identification. In this paper, we are implementing these stages. By finding the session of the user we can analyze the user behavior pattern by the time spent on a particular web page [13].

Nirali Honest [2015] et al. incorporates work on path completion by considering diverse sorts of path generated in accessing the website designed using CMS and gives a novel algorithm to form the path. Path completion is a critical and troublesome undertaking in the preprocessing stage of WUM. We shape the data preprocessing phase to achieve our goal to mine websites designed using a content management system (CMS). The data preprocessing stages include data cleaning, user identification, session identification, site structure and link details formation, path completion and event generation [14].

Saucha Diwandari [2015] et al. the outcome shows that SMO algorithm forms a better classifier model with the outcome accuracy of 95.8904% and this outcome is higher when contrasted with two other algorithms. It can be concluded that the SMO algorithm is efficient in performing classification for this case. User interaction with web sites generates a large amount of web access data stored in the web access logs. These data can be used for e-commerce to conduct an evaluation of possessing website pages as one of the efforts to understand the desires of the user. By classification techniques in WUM, we conducted an experiment to categorize a number of data obtained from the client log files in two groups, namely interest page and un-interest page by using the model page interest estimation [15].

Bhupendra Kumar Malviya [2015] et al. present that WWW is an enormous store of links and web pages. It provides a gigantic measure of data for the Internet clients. The development of the web is great as about one million pages are added every day. User's accesses are followed in web logs. Because of the immense usage of the web, the web log records are expanding at an all the more quickly rate and the range is getting to be noticeably gigantic. WUM relates mining techniques in log data to gather the performance of users, which is used in various applications like Support for the Design, E-commerce, Modified services, pre-fetching etc. WUM has three types as preprocessing, pattern detection and pattern learning. Web log data is generally noisy and confusing, thus preprocessing and pattern analysis is an essential method before mining. For learning patterns, gathering are to be constructed professionally. This paper is presents work done in the WUM. Subsequently a glance of quite a lot of functions of WUM is offered. WUM has developed into a dynamic region of study in the field of data mining because of its crucial values. This paper affords a widespread conversation of the all the stages in WUM and Problems with related works in this research area [16].

G. Dhivya [2015] et al. present that in recent years, it has been witnessed that the ever-exciting and upcoming publishing medium is the WWW. A lot of the web content material is unstructured so gathering and making feel of such data is very tedious. Web servers worldwide, generate an inconceivable measure of data on web users' browsing activities. A couple of researchers have studied these so-called web access log data to better realize and signify internet users. Knowledge can be enriched with expertise in regards to the content of visited pages and the foundation (e.g., Geographic, organizational) of the requests. The purpose of this venture is to

analyze user habits by means of mining enriched internet access log data. The a number of web usage mining methods for extracting useful aspects is discussed and hire all these methods to cluster the users of the domain to study their behaviors comprehensively. The commitments of this proposal are a data improvement that is substance and origin headquartered and a treelike perception of standard navigational arrangements. This visualization enables forward and with no problem interpretable treelike view of patterns with highlighted imperative knowledge. The results of this project can likewise be used in various purposes, together with marketing, web content advising, (re-)structuring of websites and a few other E-business methods, like recommendation and advertiser systems. It also ranks the best significant documents based on Top K query for effective and efficient data retrieval system. It filters the web documents by providing the significant content in the search engine result Page (SERP) [17].

Ying Liu [2014] et al. presents that English person names are translated into Chinese person names using the combined method-dictionary, Entropy alignment mannequin and WM. Entropy arrangements demonstrate makes utilization of the word reference of individual names and surnames, bidirectional conditional probability and transliteration similarity. Web mining makes use of rules, clue words, transliteration similarity and conditional probability. The trial comes about to appear word name dictionary combined with the entropy alignment model can achieve high precision and recall rate for large scale of parallel corpus. Web mining helps to toughen the precision of identifying translation, especially for those fallacious alignments and nonaligned names of entropy alignment mannequin [18].

LakshmanaPhaneendraMaguluri [2014] et al. present that Now-a-days, the primary focus of the search techniques in the first generation of the Web is accessing relevant documents from the Web. Although it satisfies consumer requirements, however, it is inadequate as the user mostly wishes to entry actionable expertise involving tricky relationships between two given entities. Finding such problematic relationships (sometimes called semantic associations) is especially priceless in purposes similar to National Security, Pharmacy and business Intelligence etc. For this reason the subsequent frontier is discovering primary semantic relationships between two entities reward in gigantic semantic metadata repositories. Given two substances, there exists a gigantic measure of semantic relationship between two components. Hence positioning of those affiliations is required to be in a position to discover more imperative affiliations. For this Aleman Meza et al. proposed a process involving six metrics viz. Context, subsumption, rarity, reputation, organization length and believe. To compute the total rank of the affiliations this system computes context, subsumption, rarity and status values for every aspect of the affiliations and for all of the affiliations. Nevertheless, it is apparent that, many components appear repeatedly in many affiliations consequently it isn't fundamental to compute context, subsumption, rarity and popularity values of the components every time for each affiliations rather than previously computed values may be used while computing the overall rank of the affiliations. This paper proposes an approach to reuse the previously computed values utilizing a hash data constitution hence diminish the execution time. To illustrate the adequacy of the proposed procedure, experiments have been performed on SWETO ontology. Outcome exhibit that the proposed approach is extra effective than the other current ways [19].

Ana Kovaevi [2014] et al. present that Cyber bullying has become an intensive field of research, due to its major impact on society. Most researchers analyze causes and consequences of cyberbullying, however, only a few try to improve software to reduce or stop cyber bullying, and make the Web a more secure place. In this article, current review of efforts in cyber bullying detection using web content mining techniques is presented [20].

Imre J. Rudas [2014] et al. presents that Web mining objectives to realize valuable information or skills from the web hyperlink structure, page content and utilization log. The “user-centered” philosophy of this tool is in excellent harmony with the ideas of state-of-the-art advertising, ergonomics, and studying administration. This new technique, as opposed to the ordinary “page-established” philosophy, places the users’ ambitions and intentions to the center, and

designs the offerings of the system consequently. About 60 learners of specialized teacher training partook simultaneously in processing the Educational technology and multimedia course, all their activities performed in Model learning environment were registered in a log file by the server. The processing of this log file was once full fill through the IBM SPSS net Mining for Clementine program. Here we're going to reward the first outcome exposed by using exceptional assurance in reference to the scholars' finding out endeavor, the definite pattern of the syllabus as well because the navigational possibilities. Through analysis of student behavior we acquire some beneficial expertise for path development and learning management [21].

H. Jiawei [2014] et al. presents that to improve web services, WM technology applies. We will get navigation patterns after applying web mining on web sessions those are important for web users such that proper actions can be adopted. Due to huge data in web, discovery of patterns and their analysis for further improvement in website becomes a real time necessity. In our paper calculation start from dataset comprising web sessions. For each web sessions we stream towards the route way, which shows actual traversing of user on the website. Applying the algorithm of construction of graph we will obtain in the initial graph showing each browsing session. Our goal is to enhanced mining efficiency by applying a path traversal algorithm and get surfing pattern. Our algorithm will modifies previously and one also fetches colossal dataset of web sessions, and likewise experimental results will show efficient browsing sample which can be larger for locating the user's curiosity. In modified algorithm we avert undesirable and repetition of data. Here the experimental analysis also use for enhancement of web design, target advertisement, web design improvement, customer satisfaction and efficient market analysis [22].

Rosli Omar [2014] et al. presents a review of literature containing latest works done in this field. Our aim is to provide an outline of WUM concepts relevant to pattern mining phase of WUM process. We provide reviews of pattern discovery algorithms which utilize association rules, classification and sequential patterns, and since sequential pattern mining are gaining much interest from a WUM research community extra emphasis is given to related papers [23].

CONCLUSION

In this report, we have focused three different types of web mining, namely Web Content Mining, Web Structure Mining and Web Usage Mining. WUM model is a kind of mining of server logs. WUM plays an important role in gathering, enhancing the usability of the website design, the improvement of customer's relations and improving the requirement of system performance and so on.

REFERENCES

[1]. V.Chitraa, Dr. Antony SelvdossDavamani "A Survey on Preprocessing Methods for Web Usage Data" (IJCSIS) International Journal of Computer Science and Information Security, Vol. 7, No. 3, 2010.

[2]. Dr. Sanjay Kumar Dwivedi and BhupeshRawat," A Review Paper on Data Preprocessing: A Critical Phase in Web Usage Mining Process", 2015 International Conference on Green Computing and Internet of Things (ICGCIoT) IEEE, pp 506- 510

[3]. https://www.google.co.in/search?q=personalized+content+delivery+images&rlz=1C1CHZL_enIN710IN710&source=lnms&tbm=isch&sa=X&ved=0ahUKEwirw-WaxczTAhVKp48KHbyVA4kQ_AUICigB&biw=1517&bih=735#tbm=isch&q=personalized+content+delivery+in+web+content+mining+images&imgrc=VcmVqhU2r-F7tM:

[4]. http://wwwconference.org/proceedings/www2002/refereed/504

[5]. Richa Patel, Akshay Kansara, "Web Usage Mining: A Survey on User's Navigation Pattern from Web Logs" IJSRD - International Journal for Scientific Research & Development| Vol. 2, Issue 09, 2014 | ISSN (online): 2321-0613.

[6]. P. Menaka MCA., M.Phil, A. Prathimadevi " A Survey on Web Mining and Its Techniques" © 2015, IJARCSSE

[7]. 3.Vijay Rana and DrGurdev Singh," Analysis of Web Mining Technology and

Their Impact on Semantic Web", International Conference on Innovative Applications of Computational Intelligence on Power, Energy and Controls with their Impact on Humanity (CIPECH14) 28 & 29 November 2014

[8]. .M.RekhaSundari Y.Srinivas PVGD.Prasad Reddy "A Review on Pattern Discovery Techniques of Web Usage Mining" Int. Journal of Engineering Research and Applications www.ijera.com ISSN: 2248-9622, Vol. 4, Issue 9 (Version 4), September 2014, pp.131-136.

[9]. Vivek Sharma, Mr. Sandeep Gonnade, "A Survey on Recommendation System Based on K-Nearest Neighbor Algorithm and Sentiment Analysis"

[10]. Paritosh Nagarnaik, Prof. A.Thomas" Survey on Recommendation System Methods" IEEE Sponsored 2nd international conference on eletronics and communication system (icecs 2015)

[11]. V.Anitha, Dr.P.Isakki "A Survey on Predicting User Behavior Based on Web Server Log Files in a Web Usage Mining" 2016 IEEE.

[12]. MeryemBoufim, HafidBarka, "Converting Strangers to Clients Using Web Mining Techniques", 978-1-5090-0751-6/16/$31.00 ©2016 IEEE.

[13]. G. Neelima and Dr. SireeshaRodda," Predicting user behavior through Sessions using the Web log mining", International Conference on Advances in Human Machine Interaction (HMI - 2016),March 03-05, 2016, R. L. Jalappa Institute of Technology, Doddaballapur, Bangalore, India

[14]. Nirali Honest and Dr. Atul Patel Dr. Bankim Patel "A study of Path Completion Techniques in Web Usage Mining" 2015 IEEE.

[15]. SauchaDiwandari, Adhistya Erna Permanasari, IndrianaHidayah "Performance Analysis of Naïve Bayes, PART and SMO for Classification of Page Interest in Web Usage Mining" 2015 IEEE.

[16]. Bhupendra Kumar Malviya and Jitendra Agrawal," A Study on Web Usage Mining: Theory and Applications", 2015 Fifth International Conference on Communication Systems and Network Technologies, pp 935-939.

[17]. G.Dhivya, K.Deepika, J.Kavitha and V. Nithya Kumari," ENRICHED CONTENT MINING FOR WEB APPLICATIONS", IEEE Sponsored 2nd International Conference on Innovations in Information Embedded and Communication Systems ICIIECS'15.

[18]. Ying Liu and TianJiu Xiao," Translation of English-Chinese Person Name Based on Dictionary, Bilingual Corpus and Web Mining", 2014 10th International Conference on Natural Computation IEEE, pp 818- 822.

[19]. LakshmanaPhaneendraMaguluri, M Vamsi Krishna and P S S Sridhar," A Novel Approach for Discovering Relevant Semantic Associations on Social Web Mining", 20 14 IEEE

[20]. Ana Kovaevi," Cyberbullying detection using web content mining", 22nd Telecommunications forum TELFOR 2014, November 25-27, 2014

[21]. .Imre J. Rudas and Peter Toth," Online Learning, Web Mining and Quality Assurance", 2014 International Conference on Interactive Collaborative Learning (ICL), Page 1051- 1057.

[22]. H. Jiawei, K. Micheline, Data mining concept and Techniques, second ed., Morgan Kaufmann Publishers, Elsevier inc., USA San Francisco, CA 94111 2014.

[23]. Rosli Omar, Abu Osman Md Tap, ZainatulShima Abdullah "Web Usage Mining: A Review of Recent Works"2014 IEEE.