ISSN 2321-807X
Volume 12 Number22
Journal of Advances in chemistry

Investigation of Heterogeneous Approach to Fact Invention of Web Users' Web Access Behaviour

E. Manohar, Dr. D. Shalini Punithavathani
Research Scholar, Anna University Chennai, India
manohar2k@ymail.com
Principal, Government College of Engineering, Tirunelveli, India
shalini329@gmai.com

ABSTRACT

World Wide Web consists of a huge volume of different types of data. Web mining is one of the fields of data mining wherein there are different web services and a large number of web users. Web user mining is also one of the fields of web mining. The web users' information about the web access is collected through different ways. The most common technique to collect information about the web users is through web log file. There are several other techniques available to collect web users' web access information; they are through browser agent, user authentication, web review, web rating, web ranking and tracking cookies. The web users find it difficult to retrieve their required information in time from the web because of the huge volume of unstructured and structured information which increases the complexity of the web. Web usage mining is very much important for various purposes such as organizing website, business and maintenance service, personalization of website and reducing the network bandwidth. This paper provides an analysis about the web usage mining techniques.

Indexing terms/Keywords

pattern discovery, preprocessing, web usage mining, web rating, web review, web ranking.

Academic Discipline And Sub-Disciplines

Education - Engineering

SUBJECT CLASSIFICATION

Computer Science and Engineering

TYPE (METHOD/APPROACH)

Quasi-Experimental

1. INTRODUCTION

Web consists of a huge garbage collection of information. The number of web users, the traffic in the network and the number of websites grow rapidly every day. Most of the contents in many websites are unstructured information. Due to the growth of today's technology, the web based marketing has increased rapidly. For the effective use of web pages by the web users, web personalization is very much essential. The number of web page becomes double by every month [1]. As the growth of information is rapid, the network technology also becomes faster. Nowadays, internet is commonly used by most of the people as there is a lot of information in the web and so for the large number of web users, there is a need for reducing the network bandwidth. As this era becomes faster, the supply of relevant information to the web users would also be faster. Due to the various challenges mentioned above, there is a need for using a technology to provide relevant information to the users as quickly as possible; this can be done only by the technique called web user mining. Data mining is nothing but discovery of knowledge, whereas web mining is the application of data mining. Web usage mining is the sub field of web mining where web mining is classified as web structured, web usage and web content mining [2]. Mining the web structure discovers the structure of the web pages hyperlinks in the website. Mining the web users' access behavior leads to the discovery of knowledge about the web users. Mining the web content discovers the content in a web page such as video, audio, images etc. In web usage mining, the web log file consists of three phases; they are web log preprocessing, knowledge discovery about the web users and knowledge analysis about the web users. Web log file consists of web access behavior of the web users [3]. The various methods which are used to discover the knowledge about the web users are statistical analysis, association rule mining, cluster analysis and classification technique. In knowledge analysis phase, the insignificant information is filtered. Because of web user mining, more number of web users is attracted to the website due to web personalization.

Web servers consist of various types of log files [4]. They are

- Web Access logs
- Web Agent logs
- Web Error logs
- Web Referrer logs

Web Access Logs: Web access log file consists of various collection of information such as date, time, web users' IP address and users' action on the web page.

Web Agent Logs: Web agent logs provide various collection of information such as browser, browser version and operating system of web users etc.

Web Error Logs: Error log indicates various error information such as file not found, document contain no data, configuration error, error in time, error in web user domain name and the page on which a user received the error, providing a server administrator with information on problematic and erroneous links on the server and the data written to the error log which includes stopped transmissions and information on user-interrupted transfers.

Web Referrer Logs: Web referrer logs provide information about the referrer of both the same site and other site. The web user may arrive at a website through web search engine with some particular keywords which are all obtained through web referrer logs.

Types of Web log Formats: Web server log files are in two types of format; they are Common Log file Format and Extended Log file Format. Common Log file Format consists of date (date, time, and time zone), client IP (remote host IP and DNS entry), user name (remote log name of a user), bytes transferred, server name, request (URI query) and status (http status code returned). Extended Log file Format consists of bytes sent and bytes received, server (name, IP address, and port), request (URI query and stem), requested service name, time taken for transaction to complete, version of transfer protocol used, user agent, cookie ID, and referrer.

The process of data pre-processing is known as data preparation [3]. Data preprocessing leads to improve data quality and increase mining accuracy. The input for the knowledge discovery step is the preprocessed web log file. Extracting the fields is the first step in web log preprocessing [4]. In extracting the field, the web log file in a single field is extracted into separate field. The web server mostly uses comma (,) and space (" ") character as separator for each field. Ciesielski and Lalani [5], propose a technique of how to get most of the information by using the web user IP address. Anand Sharma [6] proposes a technique to find the first 3 – last2 visited pages and the list of directories. The web log cleaning phase cleans the irrelevant and redundant data, so that the correct information can be put into the knowledge discovery phase [4]. Data cleaning reduces the log file size. Analyzing huge volume of data is a complex time consuming activity. Web log file contains information such as gif, JPEG, etc. Mohammad and Soukaena [7] represent the classification of web log file by using decision tree classifier. Youquan [8] proposed a technique to find frequent item sets employing Rough set Theory. Castellano, Fanelli, Torsello [9] deal with four modules namely data cleaning, data formatting, data filtering and data summarization. Tan P. N. and Kumar [10] represent the data cleaning by removal of outlier or irrelevant data such as robots. Kushmerick [11] represents a method which removes the advertisement in the web log. Web user identification is one of the challenging tasks in web usage mining. It is very important for web personalization where unique web user is identified [4]. Mostly, the individual web user is identified by using IP address [12]. The identification of web user is difficult because of local caches, firewalls and proxy servers. There are several other techniques to identify the web users [13]; they are proactive method which uses cookies and reactive method which uses web log file. By navigating web log file, we can identify the timing sequence of the web users. In reactive strategy, clustering technique is also used to find the navigation method [14].

Suresh R. and Padmajavalli R. [15] identify web user by using the access behaviour and by constructing the transaction. Robert Cooley, Bamshed Mobasher and Jaideep Srinivastava [16] identify web user by using cookies and authentication mechanism. Istvan K. Nagy and Csaba G. [17] identify the web user by calculating the time spent in every page. Ivancsy R. and Juhasz S. [18] use IP address to identify the web users. Morzy T., Wojcie M. and Zakrazewicz [19] use clustering method to navigate the web users' web pages access sequence to identify the web user. Spiliopoulou M., Mobasher B., Berendt. B and Nakagawa K. [20] use the cookies and navigation on web log files to identify the web users. Session is nothing but the navigation of several web pages by a web user at that particular access time period [4]. A web user has a single session or multiple sessions during a particular period. Session is considered as a particular interval of time between two visits of website by the web user [21]. The navigation-driven method consists of two methods; they are maximum forward reference transaction identification and reference length method [9]. Most of the techniques use 30 minutes as a default time out [22] and establish time out of 25.5 minutes for some observed data. Robert F., Dell P., Roman E. and Juan D. [23] obtain the session based on the program and created all session simultaneously. Jose M. Domenech and Javier L. [24] use the time oriented heuristic and reference based method to identify the session. Baoyao Zhou, Siu Cheung Hui and Alvis C. [3] set a default session time as 30 minutes to identify a session. They use time stamp. Cooley R. [16] identifies session by using the time oriented heuristics of 30 minutes. Catlegde L. and Pitkow J. [25] use the time oriented heuristics of 25.5minutes to 24 Hrs. Chitraa V., Dr. Davamani A. [26] use navigation of web user access by creating a topology in graph method to identify the session. Joining or dividing several sessions into a meaningful cluster is called transaction identification. There are various approaches of web transaction; they are reference length and transaction identification by maximal forward reference [27]. Path completion of the web pages by web users is one of the challenging tasks. In some cases, the web log information about the web users' access may not be present due to the clicking of back and forward button in the web browser by the web users and also by using proxy server [4]. In the above cases, the web links information will not be there in the web log files. So, to discover the missing path information, we use the web path completion technique to find out the missing path. Yan L., Boqin F. and Qinjiao M. [21] use referrer based method to determine the pages accessed by the web users using proxy servers and local caching. Aruna Kumari G. K. and Sudheer Shetty [28] navigate the path by using both auxiliary path completion and content path completion. Sharma A. [6] determines; the path by identifying first 3 and last 2 visited web pages by the web user and also by using a list of

directories. For the effective web log preprocessing, the following steps should be carried out effectively; they are data collection, data integration, data cleaning, data reduction, session identification and user identification [29][30].

Robert Cooley, Bamshad Mobasher and Jaidep Srivastava proposed their method for web user identification, users' session identification, page view identification, path completion and episode identification [13]. Berendt. B and Spiliopoulou M. proposed the compared time based and referrer based heuristics for path completion [31]. The preprocessing work includes identification of the web user, identification of users' sessions and path completion of the cached page.

Knowledge discovery about the web users depends on different algorithms and methods developed from various fields. The discovery of knowledge includes statistics, data mining, machine learning and pattern recognition. Several statistical analysis, association rule mining, clustering and classification and sequential pattern mining are used during different stages based on different applications and different requirements. The knowledge discovery phases based on different analysis are categorized as Association rule mining, Clustering, Classification and Sequential pattern mining [32][33][34][35].

Association rule refers to a set of pages that are associated to each other with minimum support value. If there is a maximum association, then the designer may directly link the web page with the existing web page. The association mining algorithm is an application of web usage mining where the mining is focused on the next interesting web page of the web user [35]. The various association mining algorithms are AIS by R. S. Agarwal. In this algorithm, only one item consequent association rules are generated. In SETM Algorithm by Houtsma. M and A. Swami, the new candidate item sets are generated in the same way as in AIS algorithm, but the transaction identifier of the generating transaction is saved with the candidate item set in a sequential structure. It separates candidate generation process from counting. At the end of the pass, the support count of candidate item sets is determined by aggregating the sequential structure. In Apriori algorithm by Zhao, the Candidate item sets are generated using the previous pass without considering the transactions in the database. In Apriori TID by Ceglaret, 'C' is generated of which each member has the Transaction ID of each transaction and the large item sets are present in this transaction. This set is used to count the support of each candidate item set. In FP growth Algorithm by Han & Pei, Divide and Conquer Technique is used for the association of web log data in FP growth algorithm. In RARM by DAS, Ng. and Woon, a versatile tree structure known as the Support Ordered Tree Item set structure is used to hold pre-processed transactional data for the association of web log data. In Improved AprioriAll by Wang Tong, the property of the User ID is used during every step of producing the candidate set and every step of scanning the database by which to decide whether an item in the candidate set should be put into the large set which will be used to produce next candidate set. In the mean time, in order to reduce the number of the database scanning, the new algorithm, uses the property of the Apriori algorithm which limits the size of the candidate set in time whenever it is produced. In Custom-built Apriori algorithm by Sandeep sing, the Apriori algorithm can be customized in such a way that pruning operation is performed only on the candidate item sets whose size > 2. For generation of frequent item sets of size k, only the transactions whose size >= k are considered. Markov Chain Model by Andrey Markov, uses probability theory. A Markov model is a stochastic model used to model randomly changing systems where it is assumed that future states depend only on the current state not on the events that occurred before it. Association rule hiding algorithm by Natarajan R., Dr. R. Sugumar, M. Mahendran and Anbazhagan is used to implement privacy preserving data mining. The association rule hiding algorithm for privacy preserving data mining would be efficient in providing confidentiality and improving the performance at the time when the database stores and retrieves huge volume of data. In a maximal forward reference by Chen, a maximum forward reference is defined as the longest consecutive of forward reference before the first backward reference is made. In Markov Chain by Andrey Markov, a Markov chain is a random process that undergoes transitions from one state to another on a state space. It must possess a property that is usually characterized as "memorylessness": the probability distribution of the next state depends only on the current state and not on the sequence of events that preceded it [36].

The most widely used knowledge discovery technique is the clustering method [37]. Clustering is grouping similar data for discovering the knowledge. The various types of cluster to be discovered in web usage mining are page cluster, and usage content [38]. This knowledge discovery is useful to personalize the web content of the web users. The knowledge discovery is very useful in web assistance and search engine personalization [35]. Clustering can be broadly classified into three types; they are partitioning methods, hierarchical methods and model based methods. In partitioning method, the data are divided into k groups. Various algorithms can be applied for different purposes [35]: In Expectation-Maximization by Arthur Dempster, Nan Laird, and Donald Rubin, Expectation–maximization algorithm is an iterative method for finding maximum likelihood or maximum a posteriori estimates of parameters in statistical models, where the model depends on unobserved latent variables. In Fuzzy clustering by Wolfram, Fuzzy clustering is a form of clustering in which each data point can belong to more than one cluster or partition. In Graph partitioning by Hendrickson and Leland, the graph partition problem is defined as data represented in the form of a graph, with V vertices and E edges, such that it is possible to partition G into smaller components with specific properties. For instance, a k-way partition divides the vertex set into k smaller components. In Self-Organizing Maps by Teuvo Kohonen, a self-organizing map or self-organising feature map is a type of artificial neural network that is trained using unsupervised learning to produce a low-dimensional, discretized representation of the input space of the training samples called a map. In Ant-based by Marco Dorigo, the ant colony optimization algorithm is a probabilistic technique for solving computational problems which can be reduced to finding good paths through graphs.In K-means with genetic algorithm determined by Krishna and Narasimha Murty, Genetic algorithm finds a globally optimal partition of a given data into a specified number of clusters. Page Gather by Cutting, Karger, Pedersen, and Tukey initially the system scatters the collection into a small number of document groups or clusters, and presents short summaries of them to the user. Based on these summaries, the user selects one or more of the groups for further study. The selected groups are gathered together to form a sub collection. The system then

applies clustering again to scatter the new sub collection into a small number of document groups which are again presented to the user. In each successive iteration, the groups become smaller and therefore more detailed. The web data are decomposed to create hierarchical structure of the cluster. In Hierarchies (BIRCH) algorithm by Tian Zhang, Raghu, Ramakrishnan, Miron Livny, BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) is an unsupervised data mining algorithm used to perform hierarchical clustering over particularly large data-sets. Model based methods identify the best fit between given dataset and mathematical model. There are different algorithms used in mathematical model for clustering the web users' session.

In Autoclass by Stutz and Cheeseman, Autoclass is a clustering algorithm of mixed data type which uses a Bayesian method to determine the optimal classes based on prior distribution. Self-Organizing by Juha Vesanto and Esa Alhoniemi, projects input space on prototypes of a low-dimensional regular grid that can be effectively utilized to visualize and explore properties of the data. When the number of SOM units is large to facilitate quantitative analysis of the map and the data, similar units need to be grouped. Fishers, COBWEB is an incremental algorithm which takes an object at a time to decide whether it should be accommodated in the existing cluster or added to the hierarchy as new cluster.

Classification is used to discover knowledge by classifying the information into two different classes [35]. To identify different classes, the web users' web page can be personalized based on the specific class [38]. Classification may be used to identify the web users under different categories. There are different classification algorithms for web usage mining for different purposes [34][39]. HCV by Wu, HCV receive input either through the Web pages extracted from the Web server log, or a set of keywords provided by the users. The result of the inductive process was a set of rules representing the users' interests, CDL4 by Shen, is a semi incremental learning method. In RIPPER by Cohen, RIPPER is abbreviated as Repeated Incremental Pruning to Produce Error Reduction. RIPPER is especially more efficient on large noisy datasets. C4.5 by Ross Quinlan, builds decision trees from a set of training data using the concept of information entropy. In Naive Bayesian by Edwards, Naive Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes theorem with strong independent assumptions among the features. Rough Set Theory by Zdzisław I. Pawlak, is the approximation of crisp set in terms of a pair of sets which give the lower and the upper approximation of the original set.

Knowledge discovery using sequential pattern mining identifies the occurrence of sequential events in order to determine whether there is any relevant sequence in that occurrence [40]. Discovering knowledge can be used to predict the next page the web users are going to access and also it guides the designer to personalize the advertisement to the web users. The knowledge discovery using sequential pattern mining can be determined by using two methods; they are deterministic method and stochastic method. Deterministic method holds the navigational behaviour of the web users. Stochastic method uses sequential web page access in order to predict the next visit by the web users [41]. The sequential mining using deterministic approach consists of various algorithms [42][43]. GSP Algorithm by Ramakrishnan Srikant and Rakesh Agrawal [41], is an algorithm used for sequence mining. The algorithms for solving sequence mining problems are mostly based on the apriori algorithm. One way to use the level-wise paradigm is to first discover all the frequent items in a level-wise fashion. As Wei Shen, Jianyong Wang and Jiawei Han [42], put it, Frequent pattern projected sequence pattern mining integrates the mining of frequent sequence with that of frequent patterns and uses projected sequence database to confine the growth of subsequent frame. Sequential Pattern discovery using Equivalence classes is determined by Philippe Fournier-Viger, Ted Gueniche, Souleymane Zida, Vincent S. Tseng Jozef K., Michal M. and Martin D. [43]. ERMiner search uses equivalence classes of rules having the same antecedent or consequent. Data Mining of User Navigation Patterns by Borges Mark Levene, extracts the navigation pattern from web user sessions. Markov models by Andrey Markov, is a stochastic model used to model randomly changing systems where it is assumed that future states depend only on the current state not on the events that occurred before it.

After knowledge discovery, the knowledge analysis is the final phase. Knowledge analysis phase identifies the relevant information from the knowledge discovery phase. The knowledge analysis phase drops all the knowledge which is not relevant to the particular application [44]. The pattern of the web users is identified by navigating the web users' access behaviour of the web page. Visual representation of the knowledge makes it easy to interpret.

Most of the web ranking methods are based on web structure analysis. The ranking of a web page is awarded based on the number of incoming links to the web page and number of outgoing links from the page. In [45], the ranking of a web page was calculated based on the number of link hits to that web page. Web page ranking of journal citation is reviewed using a novel method [46]. The normalization technique is used to improve the web page rank and it explains how to improve the web page rank [47]. Although there are several web ranking techniques and web rank improvement techniques, most of the web page rank is based on web page links. Neelam Tyagi and Sharma [48] weighted the rank of the page based on the number of visits of links of web pages in time duration.

Web review is the effective method to discover the web users' behaviour as in web review, the web users share their views. Bing, Bu, Chen, and Qiu [49] proposed a novel propagation based method to solve option lexicon expansion and target extraction problem simultaneously. Yi, Nasukawa, Brunescu, and Niblack [50] present sentiment analyser, which extracts sentiment about a subject online text document. Chen, Qi and Wang [51], compare the methods based on a real world review data set, but also in particular, adopted the conditional random field's model. Xianghua, Guo, and Yanyan [52] proposed an unsupervised approach to discover automatically the aspects discussed in Chinese social review and also sentiments expressed in different aspects. Hua Xu, Weiwei Yang and Jiushuo Wang [53] proposed classification and emotion component analysis on chinese micro-blog posts. Hu and Liu [54] focus on mining opinion/product features of which the reviewers have commended on. Web review gives the exact representation about the web users' behaviour. Sheng Chung Ding and Ting Jiang [55] represent an optimization based approach to extract

automatically sentiment expression for a given target from a corpus of unlabelled tweets. Wang, Tsou, Zhu. J., Zhu. M and Ma [56] proposed opinion mining approach in order to apply it to the tourism domain. Erdem Ucar, Erdinc Uzun and Pınar Tufekci [57] proposed a method for extracting review from web pages forums, blogs, newspapers, commerce and trips.

Many online sites have web rating scale to discover the knowledge about the web users. The web rating scale is widely used in e-commerce sites. Mostly the rating scale is rated in 5 point scale. The web rating scale indicates the web users' views on the particular product or a particular page. Martin Emmert and Florian Meier [58] perform analysis on physician rating websites. It has been gaining in popularity among patients who are seeking a physician. Patel, Cain, Neailey and Hooberma [59] analyze the patients' views by using web-based feedback and ratings of the general practitioners. The web rating scale is used in various applications to measure the web users' opinion about the page or a particular product.

2. PROPOSED METHODOLOGY

In this proposed methodology, we discover the knowledge about the web users by considering various techniques to discover the web users' access behaviour. Here we use four different methods to discover the web users' behaviour in the website. They are 1. Web log based web users' behaviour analysis. 2. Ranking based web users' behaviour analysis. 3. Review based web users' behaviour analysis and 4. Rating based web users' behaviour analysis.

2.1 Web Log Based Web Users' Behaviour Analysis

The web log based web users' behavior analysis is the most common and effective method to analyse the web user access behavior. In this methodology, the web log file from the web server is used for analyzing the web users. All the web users' access in the website is stored as web log file. Web users' web log file consists of 16 attributes. By using the web log file, we can identify the web user and their access behavior. This methodology of identifying the web users' behavior analysis consists of three phases. They are 1. Web log preprocessing 2. Knowledge discovery 3. Knowledge analysis.

A. Web log Pre-processing

Pre-processing is the most important phase in behaviour analysis. Before processing the web log file, we have to pre-process the web log file; so that the result of the analysis will be accurate and also we save the time in knowledge discovery. Pre-processing consists of several steps. They are 1. Data consolidation 2. Data cleaning. 3. Data transformation. 4. Data reduction.

Data consolidation

Data from different web servers are to be consolidated to a single file. Integration of web log file from multiple web servers is called data fusion. It provides a unified view of the data. Here we consolidate the data from the heterogeneous source.

Data cleaning

Data cleaning is the process of removing incorrect, incomplete, improper and duplicate data. Data cleaning makes the knowledge discovery more accurate. The size of the input log file is reduced.

Data transformation

The data is transformed in the appropriate form which is suitable for mining by aggregation. In our proposed methodology, we aggregate the data based on ip address, so that we can identify the unique web users. There is still a problem in identifying the web users because many users may have access to the website from the same ip address; this is more common access to the website through the proxy server. This problem can be resolved by using the sequential pattern mining for the particular ip based on the time sequence by which each session of the web users can be identified

Data reduction

Data reduction is the transformation of data into corrected, ordered and simplified form. Data reduction transforms the data into meaningful data. In this process, we have to remove the various attributes in the web log file which are not necessary for the knowledge discovery phase.

B. Knowledge Discovery

In the knowledge discovery phase, we use Apriori algorithm for the frequently accessing page of every unique web user by considering each session by the web users. It is an association rule learning method. Apriori algorithm uses bottom up approach where frequent page extends one page at a time, which is called candidate generation. The groups of candidates are tested against the data. The algorithm gets terminated where no successful extension is found. Here we use the breadth first search to search the next page and the hash tree to count candidate item set.

Algorithm – Web log based web users' behaviour analysis.

Input : Web log file

Output : Discovered knowledge about the web users

Procedure:

Step 1: Web fusion

Step 2: Removing incorrect, incomplete, improper and duplicate web log file.

Step 3: Data Transformation by aggregating the web log based on ip address

Step 4: Sequential pattern mining for identifying unique user based on the ip and the web user session.

Step 5: Removes various attributes in the web log files such as file extension, user agent, and browser.

Step 6: Associate the web users' behaviour by using Apriori Algorithm.

Step 7: Organize the knowledge by removing the number of page visit by the user which is less than 2.

C. Knowledge Analysis

In knowledge analysis phase, we are going to analyse the discovered knowledge. The knowledge is analysed based on different applications.

2.2 Ranking Based Web Users' Behaviour Analysis

In ranking based web users' behaviour analysis, we use the page rank concept to identify the web users' access behaviour. Most of the page ranking tools are based on the web page links. If the number of incoming link to the web page is high, then most page ranking algorithm assigns maximum rank to that page.

In our ranking based web users' behaviour analysis the visitor hit counter is a front-end tool that simply shows how many visitors have been on a website or web page. Each time a new visitor visits that website or page, the counter will increase by one more number. Whenever a web page is accessed, the script will be loaded on client side from the web server. Script will monitor the click on the web page. On the server side, the web log file is used to record the webpage id, hit count of the hyperlinks. Hit count will increment whenever hit occurs on the hyper links. The crawled information is stored in the search engine database which is used to calculate the rank value.

Calculation of page hits:

If 'p' is the page, 'PH' is the hit count of each page. The counter increments whenever the page is viewed by the web users.

$$PH_i = PH + 1 \quad (1)$$

Page rank based in hit count:

Let 'p' be the page and 'n' is the number of pages. 'PR' is the page rank based on page hits. Arrange the page in the descending order by using quick sort algorithm. Then rank the page in sequence where the page in the top of the order holds the first rank and the page at the bottom of the order holds the last rank. 'q' is the sort function

$$PR = q(PH_i) \quad (2)$$

$$PR_{First} = (\mathrm{PR} == \mathrm{MAX}) \quad (3)$$

$$PR_{Last} = (PR == MIN) \quad (4)$$

2.3 Review Based Web Users' Behaviour Analysis

In review based web users' behaviour analysis, we use the web users' review to analyse the web users' behaviour. In this method, we extract the opinion target and option word from the given comments. The relationship between the word and the target is found. Next find the synonym of the word by using semantic algorithm. After that, the opinion word is compared with the good word data set and bad word data set. If it is available in the good word list, the good word counter will be incremented; if it is in the bad word list, then the bad word counter will be incremented.

In this model, the review is to be pre-processed. In pre-processing we are going to generate the token which identifies the character and the symbol. Then prune the unwanted symbols. Then identify the stemming word.

The model assumes that all nouns/noun phrases in the sentences are opinion target candidates; all adjectives/verbs are regarded as option words. After extraction of opinion target and opinion word, the next step is to classify the opinion word. By using the semantic algorithm, we are going to identify the related meaning of the word.

Next, we have to classify the word by using KNN classification. In K nearest neighbourhood classifier, k is a Positive integer If k=1, the object is assigned to single nearest neighbourhood. By counting the number of positive review, we can measure the web user behaviour.

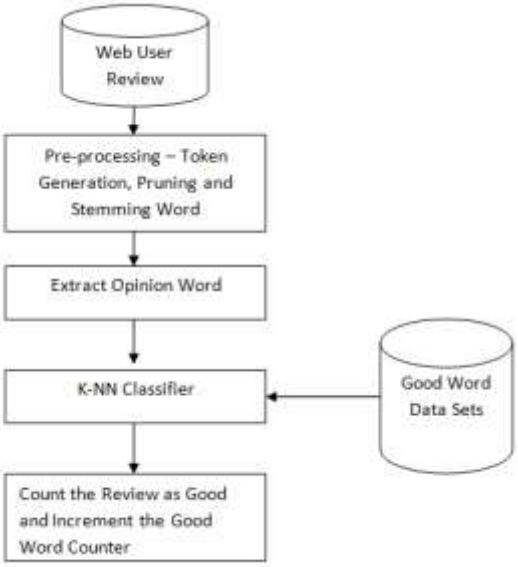

Fig 1: Architecture of web review analysis

2.4 Rating Based Web Users' Behaviour Analysis

In rating based web users' behaviour analysis, the web users' behaviour is measured based on the rating given by the web user in the 5 point scale. In rating based analysis, each web user can rate the web page in the following points; they are 1, 2, 3, 4, 5 where the value of each page is calculated by using the formula

$$PG_{Rate(i)} = \left[\left(5 * Count_5\right) + \left(4 * Count_4\right) + \left(3 * Count_3\right) + \left(2 * Count_2\right) + \left(1 * Count_1\right)\right] / N \quad (5)$$

Where 'PG_{rate}' is the web users rating of the particular page, 'i' is the page identity number where i = 1 to N, 'count5' is the count of 5 rating in the page, 'count4' is the count of 4 rating in the page, 'count3' is the count of 3 rating in the page, 'count2' is the count of 2 rating in the page and 'count1' is the count of 1 rating in the page and 'N' is the number of users. After calculating the page rating of all the pages, we have to sort the page based on the 'PG_{rate}' in descending order in the list using linear sort. The LIST == MAX is the most liked page and where the LIST == MIN is the most disliked page.

3. RESULT AND ANALYSIS

In this section, we analyse the web users' behaviour based on various analysis about the users of loadpict.com website. The study shows that knowledge discovery about the web users by using web log file is more accurate than the other two methods. While considering the system complexity, ranking based analysis based on hit count of the page is less complex to identify the web users' behaviour where personalized behaviour analysis is little complicated where it needs web log file. While considering the other two methods, review based web users' analysis makes a detailed study of the web users as the web users share their views openly. In review based web usage analysis, there are some drawbacks. Although we analyse the behaviour more accurately, all the web users do not participate in the web review; so the major drawback is a limitation in identifying the access behavior of more number of web users. The final web users' behaviour analysis method is the web rating analysis. The behaviour analysis by this method is also more accurate as the web users openly share their view with some limitation in the way of expressing in numerical values. The performance of the rating analysis methodology depends on the number of web users' participation in web page rating.

We analyse the various techniques by considering 30 web users of loadpict.com for 15 days. We analyse the web user behaviour using web log analysis, web ranking analysis, web rating analysis and web review analysis. The Table 1 below shows the measurements.

Table 1. Measurements by various methodologies

Techniques Evaluation	Web Log Based Analysis	Web Ranking Based Analysis	Web Rating Based Analysis	Web Review Based Analysis
Total Number of Web Users	30	30	30	30
Reject Count	2	0	15	20
Good Count	28	30	15	10
Planned Time	10 Sec.	10 Sec.	10 Sec.	10 Sec.
Stop Time	5.49 Sec.	9.23 Sec.	8.01 Sec.	5.41 Sec.
Run Time	4.51 Sec.	0.77 Sec.	1.99 Sec.	4.59 Sec.
Ideal Cycle Time	0.2 Sec.	0.2 Sec.	0.2 Sec.	0.2 Sec.
Quality	0.93	1.00	0.50	0.33
Availability	0.45	0.08	0.20	0.46
Performance	1.33	7.79	3.02	1.31
Effectiveness	0.56	0.60	0.30	0.20

The above result shows that various methodologies are good at different scenarios. In the above analysis, web ranking based analysis has the maximum number of participants because all the web users' session was counted. By using web ranking, all the 30 web users' behaviour was considered for the analysis.

In web log file method, the web users who do not spare the minimum threshold time on web pages are removed. In our analysis, out of 30 web users 28 web users' web log files were considered for the analysis and 2 web users' web log files were removed from the analysis during the web log pre-processing because of the insufficient information about the web users.

The number of participants in the rating system was less. Not all the participants participated in the web rating. In our analysis, only 15 out of 30 participants participated in rating the web page. So deriving all the web users' behaviour is difficult.

In web review, the number of participants is very much minimum when compared with other techniques. Here only 10 participants participated in the web review.

The result in the Table 1 shows the Quality, Availability, Performance and Effectiveness of various methods. In quality ranking based method, it holds the maximum of 100% because by using this method, we can easily retrieve all the participant information. The quality of web log based method is 93% because of not considering 2 web users for the analysis. The quality of rating based method is 50% where the quality of review based method is 33% because of the minimum number of web users' information.

The Figure 2. shows the comparison chart of web users' participation with respect to different methodology. From this analysis, we came to know that ranking based method made use of the maximum number of web users' information whereas review based method made use of minimum number of web users' information to derive the web users' behaviour. From the graph, we can conclude that the discovery of web users' behaviour analysis is more personalized by using web ranking based method and web log based analysis method.

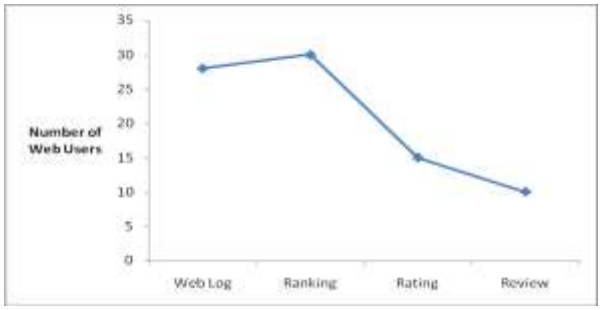

Fig 2. Web Users Participation

In Figure 3. the graphs show the execution time of various methods. Here the result of the analysis shows that the execution time of ranking based method is very less. The Figure 3. shows that the execution time of web review based method is high. The ranking based web users' behaviour analysis is very well suited for the website having a large number of audience. The web review based method is very much suited for analysing web users' behaviour of the website having a small number of web users.

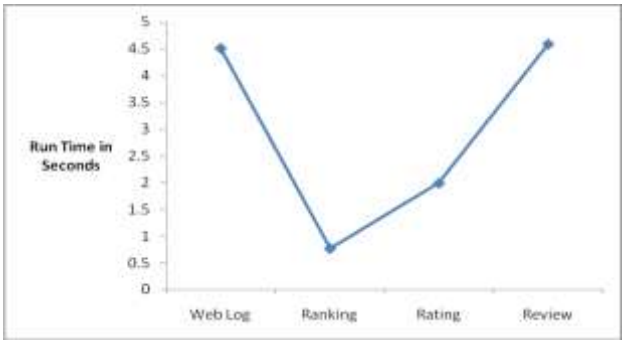

Fig 3. Execution time of various analysis methods

The quality of the various analysis methods is shown in the Figure 4. The quality of web ranking based analysis is high, as this methodology makes use of the maximum number of web users' information whereas the quality of the web review based method is low because it makes use of minimum number of web users' information.

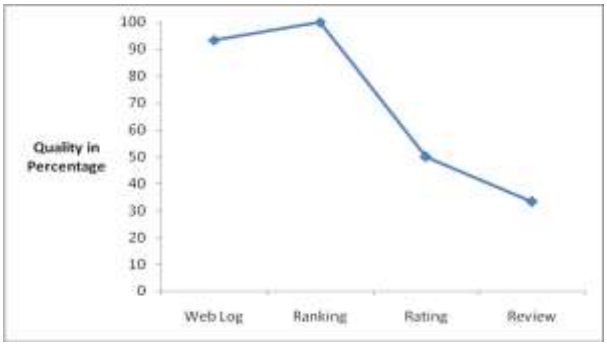

Fig 4. Quality of various analysis methods

The Figure 5. shows the availability of various methods. The availability of review based method is very much customized for the participated web users as in web review based method, the web users express their thoughts openly. But in web log based method, we discover the web users' behaviour by monitoring every activity of the web users. In ranking based method, it discovers the web users' behaviour which is more abstract. In web ranking based method, personalization of individual web users is not going to be considered where we use more generalized web users' behaviour for discovering the web users' behaviour. In rating based web users' access behaviour analysis, we are not going to analyse the web users' behaviour by personalizing the web users' activity. In web ranking based method, we are going to discover web users' web access behaviour by using the generic value evaluated by aggregation.

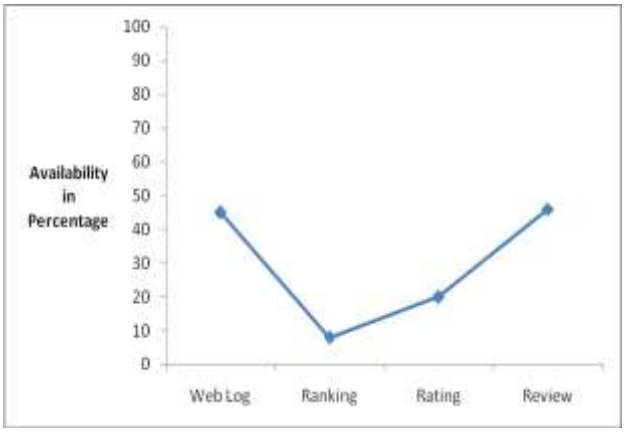

Fig 5. Availability of various analysis methods

The performance of various analysis methods is shown in the Figure 6. The performance of ranking based method is high, because the cost of ranking based method is low when compared with other methods with respect to execution time. The number of count of web users is also high when compared with other methods. The performance of web review and web log based analysis is low because in web log based analysis and web review based analysis, the cost of execution with respect to time is high, and so the overall performance is low.

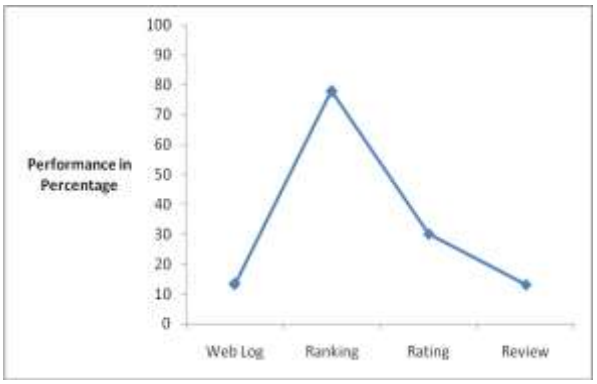

Fig 6. Performance analysis of various analysis methods

The effectiveness of various techniques is compared in the Figure 7. According to the analysis, the overall effectiveness of ranking based method and web log based method is high. The effectiveness of review based method is low because the cost of the review based method with respect to time and the number of web users' participation in review is low.

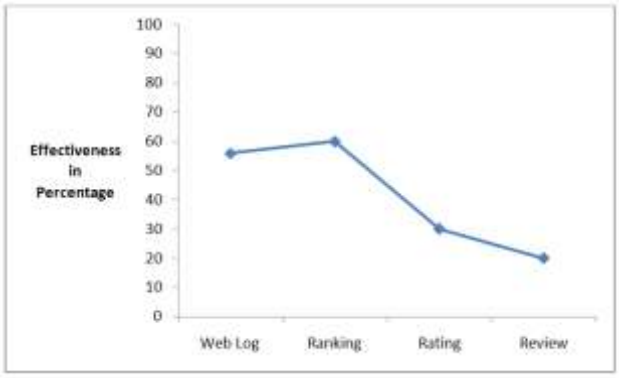

Fig 7. Effectiveness of various analysis methods

Thus by analysis, we discover the effectiveness of various techniques in discovering the web users' behaviour.

4. CONCLUSION

In this paper, we have discussed the various methods about the web users' behavior. There are various advantages and disadvantages of various algorithms and methods used. Each method is effective based on the size of the web users and the application of the various techniques. From this analysis, it is discovered that web ranking based method is more effective while analyzing the website having a large number of web users, whereas the web log based method is effective where the number of web users in the website is medium and also where we need to discover the web users' behavior analysis more personalized. The web rating based method is effective where the website is having more number of web users but in this technique, personalization of web users' analysis is not possible. Finally, in web review based method, we can analyze the web users' behavior more accurately because the web users openly express their own views but the effectiveness of the method depends on the participation of the number of web users. In future, various researches can be done by considering a large number of web users for analysis and also using various other algorithms in the above various analysis method of web users' web access behaviour.

REFERENCES

1. Duhan N., Sharma A. K., and Bhatia K. K., "Page ranking algorithms: a survey", in Proc. IEEE International Advance Computing Conference, pp. 1530-1537, 2009.doi: 10.1109/IADCC.2009.4809246
2. Facca F. M. and P. L. Lanzi, "Mining interesting knowledge from weblogs: a survey", in Proc. Data and Knowledge Engineering, vol. 53 issue 3, pp. 225-241, 2005. doi:10.1016/j.datak.2004.08.001

3. Baoyao Z., Siu C. and Alvis C., Fong M. “An Effective Approach for Periodic Web Personalization,“, in Proceedings of the IEEE/ACM International Conference on Web Intelligence. IEEE, 2006. doi: 10.1109/WI.2006.36
4. Pooja Kherwa and Jyotsna Nigam,” Data Preprocessing: A Milestone of Web Usage Mining” in International Journal of Engineering Science and Innovative Technology (IJESIT) Volume 4, Issue 2, 2015. doi: 10.1007/978-3-540-88081-3_7-Springer
5. Ciesielski V. and Lalani A., “Data mining of web access logs from an academic web site”, in Design and application of hybrid intelligent systems, pg. 1034-1043, 2003. doi : 10.1145/951440.951443
6. Sharma A.,”NY Web Usage Mining: Data Preprocessing, Pattern Discovery and Pattern Analysis on the RIT Web Data”, in Rochester Institute of Technology, 2008. doi: 10.1007/978-3-540-88081-3_7-Springer.
7. Mohammad A. and Soukaena H.,”Adding new level in KDD to make the usage mining more efficient”, National Information Technology Symposium (NITS 2006), pp. 131-140, 2006. doi : 10.1186/s40537-014-0007-7
8. Youquan H., ”Decentralized Association Rule Mining On Web using Rough Set Theory” in Journal of communication and computer ,vol. 2, no. 7, 2005. doi : 10.1.1.86.311
9. Castellano G., Fanelli A. and Torsello M., ”Log Data Preparation for Mining Web Usage Pattern”, IADIS International Conference Applied Computing, pp. 371-378, 2007. doi :10.1.1.109.4423
10. Tan. P, and Kumar, ” Discovery of Web Robot Sessions Based on their Navigational Patterns”, in Data Mining and Knowledge Discovery, vol. 6 no. 1, pp. 9-35, 2002. doi:10.1023/A:1013228602957
11. Kushmeric N., ”Learning To Remove Internet Advertisements, “in Third Annual Conf. on Autonomous Agents. 1999. doi:10.1145/301136.301186.
12. Joshila G., Maheswari V. and Dhinaharan N.,”Web Log Data Analysis and Mining” in Proc CCSIT-2011, Springer CCIS, Vol 133, pp 459-469, 2011. doi : 10.1007/978-3-642-17881-8_44
13. Cooley R., Mobasher B., Srivastava J., “Knowledge and Information System”, in Springer-Verlag, 1999. doi : 10.1007/BF03325089.
14. Chen J., Sun L., Zaiane O. and Goebel. R., “Visualizing and Discovering Web Navigational Patterns”, in Seventh International Workshop on the Web and Databases, 2004. doi : 10.1007/978-3-540-88081-3_7-Springer.
15. Suresh R. and Padmajavalli. R., ”An overview of Data Preprocessing in Data and Web Usage mining”, in IEEE International Conference, pg. 193-198, 2006. doi : 10.1109/ICDIM.2007.369352
16. Cooley R., Mobasher B. and Srivastava J.,”Web Mining: Information and Pattern Discovery on the world wide web”, in IEEE International Conference on Tools With Artificial Intelligence, pp. 558-567, 1997. doi: 10.1109/TAI.1997.632303
17. Istvan K. Nagy and Csaba G., ”User Behaviour Analysis Based On Time Spent On Web Pages: Web Mining Application in E-Commerce and E-Services”, in Studies in Computational Intelligence, 2009, Volume172/2009, pg. 117-36, doi: 10.1007/978-3-540-88081-3_7-Springer.
18. Ivancsy R., and Juhasz S., ”Analysis of Web User Identification Methods”, in World Academy of Science, Engineering and Technology, vol: 1, 2007. doi :10.1287/ijoc.15.2.171.14445.
19. Morzy T., Wojcie M., and Zakrazewicz M., ”Web Use Clustering” in International Symposium on Computer and Information Sciences, pp. 374–382, 2000. doi : 10.1145/1081706.1081733.
20. Spilopoulou M., Mobasher B., Berendt B. and Nakagawa M,.”A Framework for the Evaluation of Session Reconstruction Heuristics in Web Usage Analysis”, in INFORMS Journal On Computing, vol. 15, no. 2, 2003. doi: 10.12691/ajss-3-2-1.
21. Yan L, Boqin F. and Qinjiao M., “Research on Path Completion Technique in Web Usage Mining”, in IEEE International Symposium on Computer Science and Computational Technology, 2008. doi : 10.1109/ISCSCT.2008.151
22. Michal Munk and Martin Drlík, “Impact of Different Pre-Processing Tasks on Effective Identification of Users’ Behavioral Patterns in Web-based Educational System,” Proceedings of the International Conference on Computational Science, vol. 4, pp. 1640-1649, 2011. doi:10.1016/j.procs.2011.04.177
23. Robert F., Dell P., Roman E., and Juan D., “Web User Session Reconstruction Using Integer Programming”, in IEEE/ACM International Conference on Web Intelligence and Intelligent Agent, vol. 1, pp. 385-388, 2008. doi : 10.1109/WIIAT.2008.181
24. Jose M. Domenech and Javier L., “A Tool for Web Usage Mining” 8th International Conference on Intelligent Data Engineering and Automated Learning, pp 695-704, 2007. doi : 10.1007/978-3-540-77226-2_70
25. Catlegde L. and Pitkow J., “Characterizing browsing behaviors in the world wide Web”, in Computer Networks and ISDN systems, vol. 27, no. 6, 1995. doi :10.1016/0169-7552(95)00043-7

26. Chitraa V. and Dr. Antony Selvadoss Davamani, ”An Efficient Path Completion Technique for web log mining”, in IEEE International Conference on Computational Intelligence and Computing Research, 2010. doi : 10.1007/978-3-540-88081

27. Ismail H. and Toroslu M., ”Graph Theoretic Approach for Session Reconstruction Problem”, in Information Sciences: an International Journal, vol. 177, no. 6, pp. 1523-1529, 2007. doi : 10.1016/j.ins.2006.05.004

28. Aruna Kumari G. K. and Sudheer Shetty, ”Web Usage Mining: Web log Pre-processing and Online Visitor’s frequent Pattern Discovery”, in International Journal of Innovative Research in Computer and Communication Engineering, vol. 4, no. 4, 2016. doi : 10.15680/IJIRCCE.2016. 0404149

29. Dipa Dixit and Kiruthika M., “Preprocessing Of Web Logs”, in International Journal on Computer Science and Engineering, vol. 2 no. 7, pp. 2447-2452, 2010. doi : 10.1.1.301.8554

30. Tasawar Hussain, Sohail Asghar and Nayyer Masood,” Web usage mining: A survey on preprocessing of web log file” IEEE Xplore, 2010. doi : 10.1109/ICIET.2010.5625730

31. Berendt B. and Spiliopoulou. M., “Analyzing navigation behavior in Web sites integrating multiple information systems.” in VLDB Journal on Databases and the Web, vol. 9, no. 1, pp. 56-75, 2000. doi : 10.1007/s007780050083

32. Aldekhail M., “Application and Significance of Web Usage Mining in the 21st Century: A Literature Review”, in International Journal of Computer Theory and Engineering, vol. 8, no. 1, pp. 41-47, 2016. doi : 10.7763/IJCTE.2016.V8.1017

33. Aarti Parekh, Anjali Patel, Sonal Parmar and Prof. Vaishali Patel, " Web usage Mining : Frequent Pattern Generation using Association Rule Mining and Clustering," in International Journal of Engineering Research & Technology, vol. 4, no. 4, pp. 1243-1246, 2014. doi : 10.17577/IJERTV4IS041467

34. Khushbu Patel, Anurag Punde, Kavita Namdev, Rudra Gupta and Mohit Vyas” Detailed Study of Web Mining Approaches-A Survey” in International Journal of Engineering Sciences & Research Technology, pp. 23-30, 2015, doi : 10.1.1.91.1602. 13

35. Pierrakos D., G. Paliouras, C. Papatheodorou, and C. D. Spyropoulos, "Web usage mining as a tool for personalization: A survey," in User Modeling and User-Adapted Interaction, vol. 13, pp. 311-372, 2003. doi : 10.1023/A:1026238916441

36. Moushumi Sharma, Ajit Das and Nibedita Roy, " A Complete Survey on Association Rule Mining and Its Improvement ", in International Journal of Instrumentation, Control & Automation (IJICA), vol. 4, no. 5, pp. 9335-9341, 2016.

37. Chitraa and Thavamani A. S., "An enhanced clustering technique for web usage mining", in International Journal of Engineering Research & Technology (IJERT), vol. 1, pp. 5, 2012. doi :10.1.1.680.4122.

38. Srivastava J., Cooley R., Deshpande M., and Tan P. N., "Web usage mining: Discovery and applications of usage patterns from web data," in ACM SIGKDD, vol. 1, 2000. doi : 10.1.1.110.432.

39. Lokeshkumar R., Sindhuja R. and Dr. P. Sengottuvelan, " A Survey on Preprocessing of Web Log File in Web Usage Mining", in International Journal of Emerging Technology and Advanced Engineering, vol. 4, no. 8, 2014. DOI: 10.1109/IV.2008.40.

40. Sajid N. A, S. Zafar, and S. Asghar, "Sequential pattern finding: A survey", in Proc. International Conference on Information and Emerging Technologies (ICIET), pp. 1-6, 2010 doi: 10.1109/ICIET.2010.5625726

41. Ramakrishnan Srikant and Rakesh Agrawal, ”Mining Sequential Patterns: Generalizations and Performance Improvements” in Advances in Database Technology - EDBT '96, vol. 1057, pp. 1-17, 1996. doi: 10.1007/BFb0014140

42. Wei Shen, Jianyong Wang and Jiawei Han, ”Sequential Pattern Mining” in Frequent Pattern Mining, pp. 261-282, 2014. doi : 10.1007/978-3-319-07821-2_11

43. Philippe Fournier-Viger, Ted Gueniche, Souleymane Zida, Vincent S. Tseng Jozef K, Michal M and Martin D, ”ERMiner: Sequential Rule Mining Using Equivalence Classes”, in Advances in Intelligent Data Analysis XIII, vol. 8819, pp. 108-119, 2014. doi : 10.1007/978-3-319-12571-8_10

44. Mohamed Koutheaïr Khribi, Mohamed Jemni and Olfa Nasraoui, ” Automatic Recommendations for E-Learning Personalization Based on Web Usage Mining Techniques and Information Retrieval” in IEEE International Conference on Advanced Learning Technologies, 2008. doi: 10.1109/ICALT.2008.198

45. Bharat Bhushan Agarwal and Dr. Mahmoodul Hasan Khan, ” An Improving Web Page Ranking based on Visit of Links with Time Factor and Cursor Movement Algorithm” in International Journal of Advanced Research in Computer and Communication Engineering, vol. 5, no. 1, pp. 83-86, 2016. doi : 10.17148/IJARCCE.2016.5119

46. Jiang Li and Peter Willett, " ArticleRank: a PageRank-based alternative to numbers of citations for analysing citation networks" in Aslib Journal of Information Management, vol. 61, no. 6, pp. 605-618, 2009. doi : 10.1108/00012530911005544

47. Hema Dubey and Prof. B. N. Roy,"An Improved Page Rank Algorithm Based on Optimized Normalization Technique" in International Journal of Computer Science and Information Technology, vol. 2, no. 5 pp-2183-2188, 2011. doi : 10.1.1.228.938

48. Neelam Tyagi and Sharma," Weighted Page Rank Algorithm Based on Number of Visits of Links of Web Page" in International Journal of Soft Computing and Engineering, vol. 2, no. 3, pp. 441-446, 2012. doi: 10.1.1.458.5810

49. Bing L., Bu J., Chen C. and Qiu G., "Opinion word expansion and target extraction through double propagation" in Computer Linguistics, vol. 37, no. 1, pp-9-27, 2011. doi:10.1162/coli_a_00034

50. Jeonghee Yi, Tetsuya Nasukawa, Razvan Bunescu and Wayne Niblack, "Sentiment Analyzer: Extracting Sentiments about a Given Topic using Natural Language Processing Techniques" in IEEE International Conference on Data Mining : ICDM, pp.427, 2003. doi:10.1145/245108.245122

51. Chen. L, Qi. L and Wang. F, "Comparison of feature-level learning methods for mining online consumer reviews" in Expert System with Applications, vol. 39, no. 10, pp. 9588-9601, 2012. doi : 10.1016/j.eswa.2012.02.158

52. Xianghua. F, Guo. L, Yanyan. G and Zhiqiang W, "Multi-aspect sentiment analysis for Chinese online social review based on topic modelling and HowNet lexicon" in Knowledge Based Systems, vol. 37, pp. 186-195, 2013. doi : 10.1016/j.knosys.2012.08.003

53. Hua Xu, Weiwei Yang and Jiushuo Wang, " Hierarchical emotion classification and emotion component analysis on chinese micro-blog posts" in Expert Systems with Applications: An International Journal, vol. 42, no. 22, pp. 8745-8752, 2015. doi : 10.1016/j.eswa.2015.07.028

54. Hu. M and Liu. B, "Mining opinion features in customer reviews" in National Conference on Artificial Intelligence, pp. 755-760, 2004. doi : 10.1016/0306-4573(90)90014-S

55. Shengchung ding and Ting Jiang," Comment Target Extraction Based on Conditional Random Field & Domain Ontology" in International Conference on Asian Language Processing, pp. 189-192, 2010. doi:10.1109/IALP.2010.81

56. Wang H., Tsou B. K., Zhu J., Zhu M. and Ma M., "Aspects-based opinion polling from customer reviews" in IEEE Transactions on Affective Computing, vol. 2, no. 1, pp. 37-49, 2011. doi:10.1109/T-AFFC.2011.2

57. Erdem Ucar, Erdinc Uzun and Pınar Tufekci" A novel algorithm for extracting the user reviews from web pages" in Journal of Information Science, vol. 42, no. 5, 2016. doi: 10.1177/0165551516666446

58. Martin Emmert and Florian Meier" An Analysis of Online Evaluations on a Physician Rating Website: Evidence From a German Public Reporting Instrument" in Journal of Medical Internet Research, vol. 15, no. 8, 2016. doi:10.2196/jmir.2655

59. Patel S., Cain R., Neailey K. and Hooberman L. " Exploring Patients' Views Toward Giving Web-Based Feedback and Ratings to General Practitioners in England: A Qualitative Descriptive Study" in Journal of Medical Internet Research, vol. 18, no. 8, 2016. doi: 10.2196/jmir.5865