Asian Journal of Information Technology 15 (17): 3407-3413, 2016
ISSN: 1682-3915

Advancement in Analysis of Preprocessing and Frequent Patterns in Web Usage Mining

[1]P. Senthil Pandian, [2]K. Karthikeyan and [3]K.N. Sivabalan
[1]Department of Computer Applications, University College of Engineering,
Anna University, Regional Office, Trichy, Tamil Nadu, India
[2]Department of Master of Computer Applications, Anna University,
Regional Office, Madurai, Tamil Nadu, India
[3]Department of Computer Science and Engineering,
Sri Venkateswara College of Engineering and Technology, Chitoor, Andhara Pradesh, India

Abstract: Enormous amount of information's are gathered and viewed through world wide web by different users. The user practices their views by entering hypertext credentials by internet with a large repository of web pages and web usage mining process is essential for efficient web site management, personalization, business and support services and network traffic flow analysis, etc., web page contains images, text, videos and other multimedia and web log file holds the information of the user accesses in the websites. The log file shall have some noisy and ambiguous data which may affect the data mining process and large quantity of web traffic should be handled effectively to acquire desired information. So the log file should be preprocessed to improve the quality of data. Preprocessing consists of data cleaning and data filtering, user identification and session identification. Two sets of log files are collected and processed to obtain experimental results. This study presents a framework for user and session preprocessing and clustering with Hidden Damage Data algorithm (HDD) and also analyzes the navigational behavior of users through an enhanced Conviction Frequent Pattern Mining Algorithm (CFPMA) to identify frequent patterns in web log data. The experimental result shows that the proposed technique achieves low execution time and higher accuracy when compared with the other existing methods.

Key words: Web log file preprocessing, NBU similarity, aggregative clustering, conviction value, frequent patterns

INTRODUCTION

In today's era web services, web based information system and maintaining web log data has become very important for the providers of web services. Web log data consists of information about the user's web browsing history. Most of web log data are automatically generated by web servers. The information available on the world wide web has been an explosive growth in the last few decades. To find the relevant information easily and precisely the users want to have the effective search tools. To reduce the traffic load, the web service providers gives the way to predict the user behaviors and personalize information. To discover patterns from the web that can be applied to real world problems such as enhancing web sites, product recommendation, better understanding the visitor's behavior etc., Generally web log mining is classified into web structure mining, web content mining and web usage mining. Web usage mining exploits data mining techniques to accomplish valuable information from the navigation behavior of the world wide web users. Web structure mining discusses the hyperlink structure of the web. Web usage mining is attractive to corporations including the government agencies to classify threats and fight against terrorism. Web content mining is focused on the evolution of methods to assist users to find the web documents. Web usage mining is the process of obtaining useful data from server logs. It is the application of data mining approaches to find interesting usage patterns from the web.

To improve website utility and user satisfaction this paper proposed a new methodology for the process of web usage mining steps. For extraction of user patterns this paper proposed a complete preprocessing technique. For data preprocessing enhanced hidden damage data

Corresponding Author: P. Senthil Pandian, Department of Computer Applications, University College of Engineering, Anna University, Regional Office, Trichy, Tamil Nadu, India

algorithm is proposed. Hyper text transfer protocol and hashset techniques are used in proposed algorithm. After data cleaning and filtering, identification of user based on IP address and identification of session based on user time, time interval and user session is proposed. The derived sessions are finally clustered using a web session clustering and user cluster. Both types of clustering is used for web personalization to derive usage profiles. Finally aggregative clustering is developed by combining user and session cluster. Pattern extraction process extracts interesting patterns from web logs. Pattern discovery draws upon algorithms and methods developed from various fields such as data mining, statistics, machine learning and pattern recognition. The frequent pattern discovery approaches are applied on raw data in this phase. Pattern analysis is the last phase in the web usage mining process. In this study, a framework is designed to extract aggregative clustering and frequent item sets from the web log data is proposed by using advanced conviction frequent pattern mining algorithm. The rest of the study is organized.

MATERIALS AND METHODS

Related work: Bianco *et al.* (2005) presented an analysis of web usage mining in the website OrOliveSur.com. This paper described the set of phases carried out including data preprocessing, data collection, extraction and analysis of knowledge. By using unsupervised and supervised data mining algorithms through descriptive tasks such as clustering, association and subgroup discovery and the knowledge are extracted. The results were discussed to provide some guidelines for improving website utility and user satisfaction. In conventional web usage mining semantic information about the web page content does not take part in the pattern generation process. To improve and ensure the quality of mined models for existing process mining approaches (Ly *et al*., 2012; Mishra and Choubey, 2012) developed a data transformation and preprocessing techniques steps. The concept of semantic log purging based on domain specific constraints were proposed. The feasibility of the approach was demonstrated based on a case study in higher education domain. Mishra *et al.* (2013) and Raju and Rajimol (2014) proposed the process of web usage mining can be applied in e-learning systems in order to anticipate the marks will obtain in the final exam of a course. A specific model for mining tool was developed to the use of experts in data mining and also for newcomers like instructors and courseware authors. By applying the pattern recognition techniques for web log data (Gupta and Gupta, 2011). Thakare and Gawali (2010) analyzed the web usage mining. Pattern recognition is defined as the act of taking in raw data and making an action based on the category of the pattern. Reddy *et al.* (2014) and Maheswari and Sumathi (2014) analyzed the identification of web usage patterns based on the user's interest or choice, thereby creating an intelligent semantic-based web usage mining technique. Taherizadeh and Moghadam (2009) and Yun and Ryu (2011) presented a technique to combine web content mining into web usage mining. To detect useful information and association rules about user's behaviors, the textual content of web pages is collected through extraction of frequent word sequences, which are combined with web server log files. Thakare and Gawali (2010) and Yu *et al.* (2012) proposed an effective and complete preprocessing of access stream before the actual mining process can be performed. To make actionable data source, the log file from different sources undergoes different preprocessing phases. Maheswari and Sumathi (2014) presented the algorithms to combine the log files from different servers, clean the incorporate web log file, identify the users and to develop the sessions for each user. In web log data Yu *et al.* (2012) proposed web sequential patterns using the gap-constrained method. By removing irrelevant or redundant items, collecting the similar users and reconstructing the web log data into a set of tuples constrained by visiting time, pre-process of the raw web log data was introduced. Raiyani and Jain (2012) and Raiyani *et al.* (2012) using the gap-BIDE algorithm in web log data with a less support threshold and gap constraints, web access patterns which were closed sequential patterns with gap constraints were developed.

The effect of semantic data on the patterns generated for web usage mining was investigated by Tony and Saravanan (2015). A framework was developed to integrate the semantic data into web navigation pattern generation process. Mishra and Choubey (2012) presented the frequent navigational patterns consisted of ontology instances instead of web page addresses. An evaluation mechanism involving web page recommendation measured the quality of generated patterns. The experimental results showed that the usage of semantic data in navigation pattern generation enhanced pattern quality and generated accurate recommendations. The weighted frequent pattern mining technique was used to determine frequent patterns by considering the weights of patterns. The weighted supports of patterns were matched to prune weighted infrequent patterns. Yun and Ryu (2011) proposed robust concept of mining exact weighted frequent patterns.

An efficient Hierarchical Frequent Pattern Analysis (HFPA) approach by Sudhamathy and Venkateswaran (2012) mined association rules from web logs by utilizing a normal Apriori algorithm. The interestingness measures were used to sort the discovered association rules after the application of pruning method. The rules that were ranked highly according to the interestingness measures were valuable to the web site administrator. Bhattacharya *et al.* (2013) presented the comparative analysis of Apriori algorithm and frequent pattern algorithm for frequent pattern mining in web log data. Apriori was the simplest algorithm used for mining frequent patterns from the transaction database. But the main disadvantage of the Apriori algorithm was that the generation of candidate set was costly, particularly if there is an existence of a large number of patterns or long patterns. Large item set property was used by the Apriori algorithm which was easy to implement but repeatedly performed database scan. Apriori also consumed more time to scan the large frequent patterns. Raiyani and Jain approach the web application could be a real website that contained the challenging aspects of real-life web usage mining, including external data describing the ontology of the web content.

An effective technique for frequent pattern mining using web logs for web usage mining was explained by Raju and Rajimol (2014). The approach was known as Intelligent Frequent Pattern Analysis. In this approach, the method was applied to mine association rules from web logs by utilizing a normal Apriori algorithm but with few modifications for enhancing the interestingness of the generated rules. Before the association rules were mined, the data was classified with fuzzy clustering which was then optimized through genetic algorithm. An improved web personalization approach discovered by Carmona *et al* . (2012) on user interested directories.

The method achieved lesser memory requirement and better processing time when compared with the non-weighted access pattern mining approaches. Joel *et al.* (2014) given two novel tree structures, incremental weighted Frequent Pattern tree based on weight ascending order and incremental WFP tree based on frequency descending order. They were efficient for interactive and incremental WFP mining to use the previous mining results and current tree structure when a minimum support threshold was altered or a database was updated. A parallel, distributed algorithm by Lin *et al.* (2011) was used to discover relational frequency patterns from very large datasets. A comparative analysis of association rule mining algorithms such as AIS, SETM and A priori and the AIS algorithm consisted of two phases. The generation of frequent item sets was performed in the first phase. The confident and frequent association rules were generated in the second phase. Less number of candidate item sets was produced for testing in every database pass by the Apriori algorithm.

Proposed technology: The flow of the proposed work is shown in Fig. 1.

Web log file: Web server log file records information about each user. Whenever a user hits a page web server automatically collects the log data. The log file contains an accurate navigational behaviour of users such as name, IP address, date, time, bytes transferred, access request. Whenever a user requests a resource from that particular site, corresponding to an HTTP request, web server writes information in a web log file. It gives significant information such as which pages were requested in a website, number of bytes sent to the user from the server and type of error occurs. The log file is generally used for debugging purpose with range from 1KB to 100MB. A sample web log file is given as algorithm:

198.168.0.93--[19/Sep/2014:12:22:37-0502]"GET/jobs/HTTP/1.1"20014159 "http://www.google.com/search?q=cluster+in+data+mining and hl=en andlr=and start =28and sa=R" "Mozilla/4.0(compatible; MSIE6.0;WindowsNT 5.1;SV1;.NETCLR 1.1.4322) "198.168.0.93"

is an IP address that can be converted to host name. refers the name of the remote user and login to the remote user, respectively. Both are usually omitted and replaced by a dash. [19/Sep/2014:12:22:37-0502] refers date, month, year, hour, minute, second and time zone. "GET/jobs/HTTP/1.1" refers the method, URL, HTTP protocol and its version. When a browser requests a service from a web server it returns HTTP status code in response to the request. In the proposed work Enhanced Hidden Damage Data algorithm is implemented for preprocessing. It can perform data cleaning and data filtering. The proposed workflow is shown in Fig. 1.

Enhanced hidden damage data algorithm: HTTP codes and Hashset techniques are used in HDD for preprocessing. HTTP codes are used to refer different status of each URL request. HTTP code error 4XX is mainly concentrated in the proposed work. This HTTP code helps to identify bad request (400), Unauthorized (401), Forbidden (403) and Not found (404). Hashset technique is used to give constant time, high output performance and optimize memory usage. During data cleaning unnecessary data and noisy data are removed.

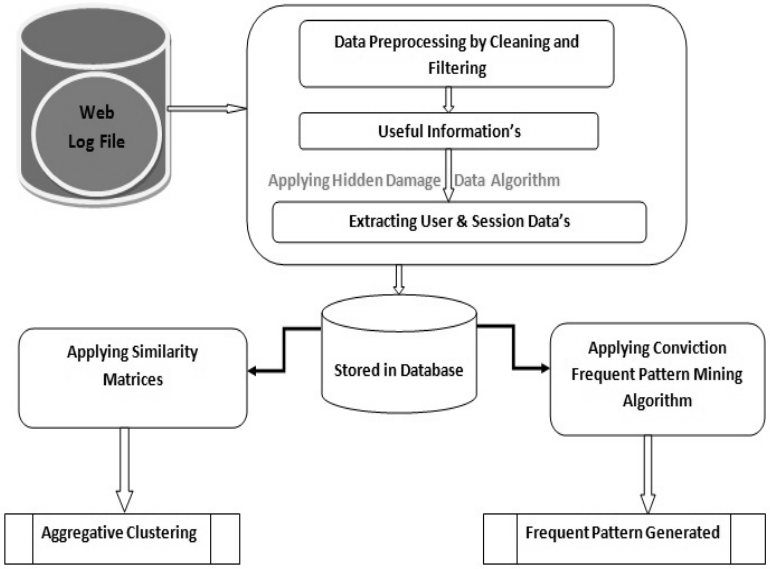

Fig.1: Formation of aggregative clustering and frequent pattern by using enhanced HDD and advanced CFPMA

As the unwanted data like irrelevant and redundant data are exactly predicted and removed, the amount of data for the knowledge extraction process is greatly reduced. So, dimensionality reduction is implemented for improving the accuracy and efficiency of further data processing is shown as.

HDD algorithm for preprocessing:

```
Start
HDD (File Input, Reader r, Writer w, File output, File corrupted)
Int corrupted Count = 0;
Int i = 0
Booleand duplicateData=false; r=ReadFile (Input);
word=StringTokenizer(r, “);
while(true)
     while(hasMoreTokens)       words[i++]=word{R};
                    I+R
  Hash Set. add(words)
End while
If (Hash Set.contains(“0”) || HashSet.contains(“-”)
   corruptedCount++; w=WriteFile(corrupted)
   corrupted.write(words)
else
    w = WriteFile(output)
    output.write(words);
If (Hash Set.size == 1)
    duplicate Data = true
return output.
```

Advanced aggregative clustering: After HTTP log files have been cleaned in data preprocessing, identification of users is done. It identifies an individual user by using their IP address. Two consecutive entries of the user's IP address are compared. If the IP address is same, operating system and the user's browser are verified. If both are same, both records are considered from the same user. Invalid user identification is removed before clustering is handled. Clustering is a process of aggregate the similar session together. It is used to pursuit hidden patterns that exist in datasets. By using the similarity metrics user clustering is applied. Based on user time, time interval and user session, session identification are implemented. The set of pages visited by a specific user at a specific time is known as session time. Depends on stay time on pages the difference between two timestamps is calculated for the time interval. A set of pages visited by the same user within the duration of one particular visit to a website is called user session. During a period a user may have a single or multiple sessions. Time based similarity metrics are calculated for session cluster. Finally, aggregative clustering is developed by combining user cluster and session cluster.

Well defined clusters with similar and dissimilar clusters are obtained in the result of aggregate clustering. After the identification of user and session, two kinds of clustering are applied. They are user based clustering and session based clustering. The user similarity metrics, Network Based User similarity (NBU similarity) is proposed. The clustering of the users depending on the network ID found in the IP or the hostname of log data is performed by using the NBU similarity. After the clustering of users, session based clustering is performed based on the time interval of the users browsed and placed according to the networked users. Aggregative clustering method is applied to combine the set of data by using set of conditions. Two constraints, user based and session based conditions are combined to provide effective clusters for the web data. So, in aggregative clustering technique, the acquired user and session cluster are combined. The similarity and dissimilarity of both the user and session based clusters are considered.

The efficiency of the dat is enhanced by reducing the irregular patterns. The time optimization is obtained by generating the aggregative results of both user and respective session time intervals.

Conviction Frequent Pattern Mining Algorithm (CFPMA): The lift and conviction value are computed for each and every item in the dataset. The lift of a rule is defined as the ratio of the observed support to that expected if X and Y are independent. It is the measure of the performance of a targeting model (association rule) at predicting or classifying cases as having an enhanced response. i.e:

$$\text{Lift}\left(X\_Y = \frac{\text{supp}(XY)+R}{\text{supp}(X)\text{supp}(Y)+R}\right) \quad (1)$$

The conviction of a rule is the ratio of expected frequency that X occurs without Y if X and Y are independent divided by the observed frequency of inaccurate predictions, i.e.:

$$\text{con}(X \rightarrow Y) = \frac{1-\text{supp}(Y)+R}{1-\text{conf}(XY)+R} \quad (2)$$

The minimum conviction value is set as a threshold limit to determine frequent patterns. The candidate itemsets with minimum conviction are collected. These itemsets are known as frequent patterns. The steps involved in frequent pattern generation process are shown in algorithm 1.

Algoritm: CFPMA for identifying frequent itemsets:

L_1= {transaction items};
for (k= 2; L_{k-1} !=Ø; k++) do begin
C_k= candidates generated from L_{k-1} L_{k-1+R} x Lk-1+R
If (C_k.sup>=min(sup)+R) then
  L_k = candidates in C_k with min_conv
end
return $U_k L_k$

Here, L_1 denotes the set of initial dataset records. C_k represents the candidate item sets that are acquired by mining least byte records by using minimum weight value. The extracted candidate item set is processed to mine frequent patterns. Before the computation of confidence, lift and conviction values, it is checked whether the support count for each item in the candidate item set (C_k) is greater than the minimum support count value. The candidates in C_k with minimum conviction value are stored in L_k. So, L_{k-1+R} is the collection of frequent patterns.

RESULTS AND DISCUSSION

Performance analysis: This chapter gives a schematic and a short description of the usage data mined from the processed log file and pattern generated. The web log data examined for evaluation is collected from the NASA Kennedy Space Center WWW server in Florida. The logs are in an ASCII file with single line per request. Two sets of web log were taken for result analysis. These two traces contain two months' worth of all HTTP requests to NASA Kennedy space center. The first log was collected from July 1,1995,00:00:00, through July 31, 1995, 23:59:59, a total of 31 day. The second log was collected from 00:00:00 August 1, 1995 through 23:59:59 August 7, 1995, a total of 7 day. A prior summary is generated as soon as the web log data are loaded into the preprocessor. The two datasets consist of total number of requests of the analyzed log file, number of corrupted or failed requests, number of satisfied requests, volume of transferred bytes, etc. After preprocessing is done then by using HDD algorithm the result obtained is shown in Table 1.

Table 1: Results of HDD algorithm

Data analysis	Results in count
Total no of records in web log	368
Data analysis	4
Time analysis	221
Time zoon analysis	2
Information analysis	196
Reply code analysis	6

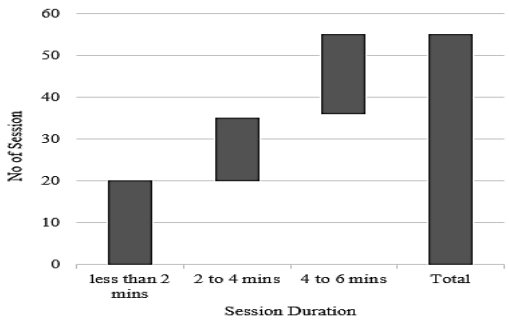

Fig. 2: Comparison between user sessions for first set of log file

A session of the user is created as long as the particular user is related to the website. Most of the time, default session time-out was taken as 30 min time-out. A session would be analyzed by a user logging into a computer, performing work and then logging off. Figure 2 shows the comparison graph between user sessions for first set of log file.

Finally, aggregate cluster is finding out by combining user and session cluster. Similar and dissimilar cluster is retrieved in the result of aggregate clustering. Figure 3 shows the aggregation of user and session cluster. For same network with same timing 15 records are obtained. For a same network with different timing 33 records are obtained.

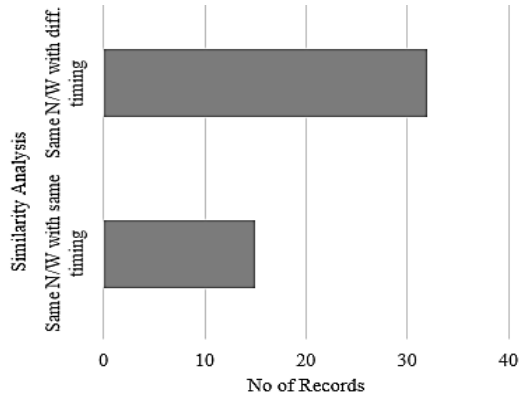

Fig. 3: Aggregation of user and session cluster

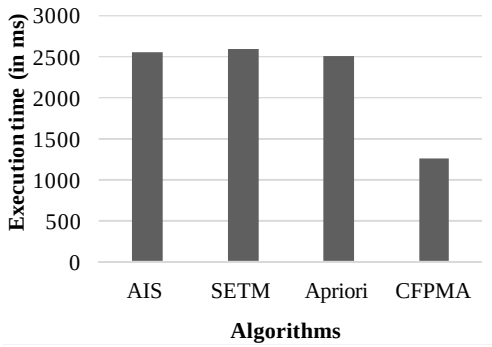

Fig. 4: Comparison of execution time consumption

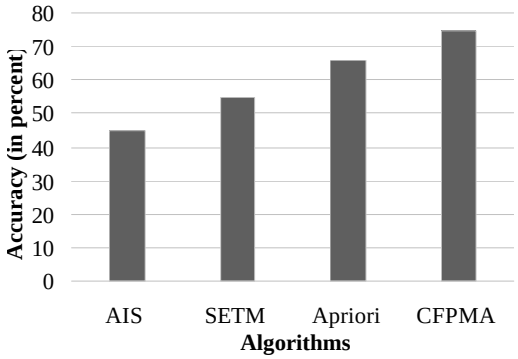

Fig. 5: Comparison of accuracy (in percentage)

The execution time consumption of the proposed approach is compared with the existing algorithms such as AIS, SETM and Apriori algorithm. With the proposed CFPMA approaches consumes less execution time and achieves higher accuracy when compared with the existing methods and the results are shown in Fig. 4 and 5.

CONCLUSION

An essential task in each data mining utilization is the formulation of an appropriate target data set to which data mining and algorithms can be applied. This study focused on web log file format, preprocessing, clustering and pattern generation techniques. Data preprocessing is implemented by using Enhanced HDD algorithms to filter and organize appropriate information. This study also deals with Advanced Conviction Frequent Pattern Mining Algorithm to extract frequent patterns from the web log data. In the existing approaches, frequent pattern mining is performed based on the support count threshold limit so that, accuracy is low and the execution time is high. But in the proposed advanced CFPMA technique, the frequent patterns are acquired based on the conviction threshold value. The experimental results gives that time taken for enhanced HDD algorithm is reduced when compared to the existing method and the proposed advanced CFPMA technique consumes less execution time and higher accuracy than other existing approaches.

REFERENCES

Bhattacharya, S., S. Rungta and N. Kar, 2013. Intelligent frequent pattern analysis in web mining. Int. J. Digital Appl. Contemporary Res., 2: 1-6.

Bianco, A., G. Mardente, M. Mellia, M. Munafo and L. Muscariello, 2005. Web user session characterization via clustering techniques. Proceedings of the IEEE Conference on Global Telecommunications Conference (GLOBECOM'05), November 28-December 2, 2005, IEEE, Turin, Italy, ISBN:0-7803-9414-3, pp: 6-6.

Carmona, C.J., G.S. Ramirez, F. Torres, E. Bernal and D.M.J. Jesus *et al.*, 2012. Web usage mining to improve the design of an e-commerce website: Or Olive Sur.com. Expert Syst. Appl., 39: 11243-11249.

Gupta, R. and P. Gupta, 2011. Fast processing of web usage mining with customized web log preprocessing and modified frequent pattern tree. Int. J. Sci. Commun. Networks, 1: 277-279.

Joel, M.R., M.V. Srinath and A. Adhiselvam, 2014. An efficient web personalization approach to discover user interested directories. J. Emerging Technol. Web Intell., 6: 142-148.

Lin, K., I. Liao and Z. Chen, 2011. An improved frequent pattern growth method for mining association rules. Expert Syst. Appl., 38: 5154-5161.

Ly, L.T., C. Indiono, J. Mangler and M.S. Rinderle, 2012. Data Transformation and Semantic Log Purging for Process Mining. In: Advanced Information Systems Engineering, Jolita, R., X. Franch, S. Brinkkemper and S. Wrycza (Eds.). Springer, Berlin, Germany, pp: 238-253.

Maheswari, B.U. and P. Sumathi, 2014. A new clustering and preprocessing for web log mining. Proceedings of the world Congress on Computing and Communication Technologies, February 27-March 1, 2014, Trichirappalli, pp: 25-29.

Mishra, A.K., M.K. Mishra, V. Chaturvedi, S.K. Gupta and J. Singh, 2013. Web usage mining using self organized map. Int. J. Adv. Res. Comput. Sci. Software Eng., 3: 532-539.

Mishra, M.R. and M.A. Choubey, 2012. Discovery of frequent patterns from web log data by using FP-growth algorithm for web usage mining. Int. J. Adv. Res. Comput. Sci. Software Eng., 2: 311-318.

Raiyani, S.A. and S. Jain, 2012. Enhance preprocessing technique distinct user identification using web log usage data. Int. J. Comput. Sci. Commun. Networks, 2: 526-530.

Raiyani, S.A., S. Jain and A.G. Raiyani, 2012. Advanced proprocessing using distinct user identification in web log usage data. Int. J. Adv. Res. Comput. Commun. 1: 418-422.

Raju, G.K. and A.N. Rajimol, 2014. A novel weighted support method for access pattern mining. Int. Arab J. E. Technol., 3: 201-209.

Reddy, K.S., G.P.S. Varma and M.K. Reddy, 2014. An effective preprocessing method for web usage mining. Int. J. Comput. Theor. Eng., 6: 412-415.

Sudhamathy, G. and C.J. Venkateswaran, 2012. An efficient hierarchical frequent pattern analysis approach for web usage mining. Int. J. Comput. Appl., 43: 1-7.

Taherizadeh, S. and N. Moghadam, 2009. Integrating web content mining into web usage mining for finding patterns and predicting users' behaviors. Int. J. Inf. Sci. Manage., 7: 51-65.

Thakare, S.B. and S.Z. Gawali, 2010. A effective and complete preprocessing for Web Usage Mining. Int. J. Comput. Sci. Eng., 2: 848-851.

Tony, A.R. and D. Saravanan, 2015. Text taxonomy using data mining clustering system. Asian J. Inf. Technol., 14: 97-104.

Yu, X., M. Li, D.G. Lee, K.D. Kim and K.H. Ryu, 2012. Application of Closed Gap-Constrained Sequential Pattern Mining in Web Log Data. In: Advances in Control and Communication, Dehuai, Z. (Ed.). Springer, Berlin, Germany, ISBN:978-3-642-26006-3, pp: 649-656.

Yun, U. and K.H. Ryu, 2011. Approximate weighted frequent pattern mining with/without noisy environments. Knowledge-Based Syst., 24: 73-82.