Journal of Computer Science 7 (5): 683-689, 2011
ISSN 1549-3636

A Novel Technique for Web Log mining with Better Data Cleaning and Transaction Identification

[1]J. Vellingiri and [2]S. Chenthur Pandian
[1]Department of CSE, Kongunadu College of Engineering and Technology,
[2]Mahalingam College of Engineering and Technology,
Pollachi, 642 003, TamilNadu, India

Abstract: Problem statement: In the internet era web sites on the internet are useful source of information for almost every activity. So there is a rapid development of World Wide Web in its volume of traffic and the size and complexity of web sites. Web mining is the application of data mining, artificial intelligence, chart technology and so on to the web data and traces user's visiting behaviors and extracts their interests using patterns. Because of its direct application in e-commerce, Web analytics, e-learning, information retrieval, web mining has become one of the important areas in computer and information science. There are several techniques like web usage mining exists. But all processes its own disadvantages. This study focuses on providing techniques for better data cleaning and transaction identification from the web log. **Approach:** Log data is usually noisy and ambiguous and preprocessing is an important process for efficient mining process. In the preprocessing, the data cleaning process includes removal of records of graphics, videos and the format information, the records with the failed HTTP status code and robots cleaning. Sessions are reconstructed and paths are completed by appending missing pages in preprocessing. And also the transactions which depict the behavior of users are constructed accurately in preprocessing by calculating the Reference Lengths of user access by considering byte rate. **Results:** When the number of records is considered, for example, for 1000 record, only 350 records are resulted using data cleaning. When the execution time is considered, the initial log take s119 seconds for execution, whereas, only 52 seconds are required by proposed technique. **Conclusion:** The experimental results show the performance of the proposed algorithm and comparatively it gives the good results for web usage mining compared to existing approaches.

Key words: Data cleaning, path completion, data preprocessing, existing data, intelligent algorithm, irrelevant data, web usage mining, fuzzy clustering, Web Robot (WR), external data, log file, reference length, transactions identification

INTRODUCTION

Recently, millions of electronic data are included on hundreds of millions data that are previously on-line today. With this significant increase of existing data on the Internet and because of its fast and disordered growth, the World Wide Web has evolved into a network of data with no proper organizational structure. Guessing the users' interests for improving the usability of web or so called personalization has turn out to be very essential and difficult in this situation.

Generally, three kinds of information have to be handled in a web site: content, structure and log data. The usage of the data mining process to these dissimilar data sets is based on the three different research directions in the area of web mining (Aziz *et al*., 2011): web content mining, web structure mining and web usage mining (Maratea and Petrosino, 2009; Jalali *et al*., 2008). Web usage mining (Liu and Liu, 2010; Chen *et al*., 2004; Wu *et al*., 1998) consists of three main steps:

- Data preprocessing
- Knowledge extraction
- Analysis of extracted results

The raw data is pretreated to get reliable sessions for efficient mining by using preprocessing. This includes:

Corresponding Author: J. Vellingiri, Department of CSE, Kongunadu College of Engineering and Technology, Thottiam, 621 215, TamilNadu, India

- Removal of records of graphics, videos and the format information
- Removal of records with the failed HTTP status code
- Robots cleaning

User identification is the process of associating page references with same IP address with different users. Session identification is breaking of a user's page references into user sessions. Path completion (Panich, 2010) is used to fill missing page references in a session. Classifications of transactions are used to know the users interest ad navigational behavior. The second step in web usage mining (Labroche *et al*., 2007; Liu and Liu, 2010) is knowledge extraction in which data mining algorithms like association rule mining techniques, clustering, classification are applied in preprocessed data. The third step is pattern analysis in which tools are provided to facilitate the transformation of information into knowledge. This study focuses on path completion process which is used to append lost pages and construction of transactions in preprocessing stage. In this study a referrer-based method is proposed to efficiently construct the reliable transactions in data preprocessing.

Related work: This section provides some of the existing techniques for web log mining. (Hussain *et al*., 2010).

The discovery of the users' navigational patterns using SOM is proposed by Etminani *et al*. (2009). Zhang *et al*. (2009) presented a Web usage mining (Chang-bin, 2010) technique based on fuzzy clustering in Identifying Target Group. Nina *et al*. (2009) suggests a complete idea for the pattern discovery of Web usage mining. Wu *et al*. (2010) given a Web Usage Mining technique based on the sequences of clicking patterns in a grid computing environment. The author discovers the usage of MSCP in a distributed grid computing surroundings and expresses its effectiveness by empirical cases. Aghabozorgi and Wah (2009) proposed the usage of incremental fuzzy clustering to Web Usage Mining. Rough set based feature selection for web usage mining is proposed by (Inbarani *et al*., 2007). Jalali *et al*. (2008) put forth a web usage mining technique based on LCS algorithm for online predicting recommendation systems. For providing the online prediction effectively, Shinde and Kulkarni (2008) provides a architecture for online recommendation for predicting in Web Usage Mining System. Exploration on web usage mining and its application was provided by Dong (2009). Huiying and Wei (2004) proposed an intelligent algorithm of data pre-processing in Web usage mining.

The usage interest on the web pages in various sessions was partitioned into clusters such that sessions with "similar" interest were placed in the same cluster using expectation maximization clustering technique was proposed.

Zhang *et al*. (2009) given an intelligent algorithm of data pre-processing in Web usage mining. Nasraoui *et al*. (2008) provides a whole framework and findings in mining Web usage navigation from Web log files of a genuine Web site which has every challenging characteristics of real-life Web usage mining, together with evolving user profiles and external data describing an ontology of the Web content. Hogo *et al*., proposed the temporal Web usage mining of Web users on single educational Web site with the help of the adapted Kohonen SOM based on rough set properties. A development of data preprocessing technique for Web usage mining and the information of algorithm for path completion are provided by Li *et al*. (2008).

Baraglia and Palmerini (2002) proposed a Web Usage Mining (WUM) system, called SUGGEST, which continuously creates the suggested connections to Web pages of probable importance for a user. Lee and Fu (2008) put forth a Web Usage Mining technique based on clustering of browsing characteristics. The approaches adopt a divide-and conquer pattern-growth principle is proposed. Filtering events using clustering in heterogeneous security logs is proposed. Mining web navigation profiles for recommendation system is suggested.

MATERIAL AND METHODS

Web log data preprocessing is a complex process and takes 80% of total mining process. Log data is pretreated (cleaning) to get reliable data. There are four steps in preprocessing of log data.

Data cleaning: The process of data cleaning is removal of outliers or irrelevant data. Analyzing the huge amounts of records in server logs is a cumbersome activity. So initial cleaning is necessary. If a user requests a specific page from server entries like gif, JPEG, are also downloaded which are not useful for further analysis are eliminated. The records with failed status code are also eliminated from logs. Automated programs like web robots, spiders and crawlers are also to be removed from log files. Thus removal process in the experiment includes

The records of graphics, videos and the format information: The records have filename extension of GIF, JPEG, CSS and so on, which can be found in the URI field of the every record, can be removed. This extension files are not actually the user interested web page, rather it is just the documents embedded in the web page. So it is not necessary to include in identifying

the user interested web pages. This cleaning process helps in discarding unnecessary evaluation and also helps in fast identification of user interested patterns.

The records with the failed HTTP status code: The HTTP status code is then considered in the next process for cleaning. By examining the status field of every record in the web access log, the records with status codes over 299 or under 200 are removed. This cleaning process will further reduce the evaluation time for determining the used interested patterns.

Method field: It should be pointed out that different from most other researches, records having value of POST or HEAD in Method field are reserved in present study for acquiring more accurate referrer information.

Robots cleaning: Web Robot (WR) (also called spider or bot) is a software tool that periodically scans a web site to extract its content. Web robots automatically follow all the hyperlinks from a web page. Search engines (Yamin and Ramayah, 2011), such as Google, periodically use WRs to gather all the pages from a web site in order to update their search indexes. The number of requests from one WR may be equal to the number of the web site's URIs. If the web site does not attract many visitors, the number of requests coming from all the WRs that have visited the site might exceed that of human-generated requests.

Eliminating WR-generated log entries not only simplifies the mining task that will follow, but it also removes uninteresting sessions from the log file. Usually, a WR has a breadth (or depth) first search strategy and follows all the links from a web page. Therefore, a WR will generate a huge number of requests on a web site. Moreover, the requests of a WR are out of the analysis scope, as the analyst is interested in discovering knowledge about users' behavior.

Most of the Web robots identify themselves by using the user agent field from the log file. Several databases referencing the known robots are maintained [Kos, ABC]. However, these databases are not exhaustive as each day new WRs appear or are being renamed, making the WR identification task more difficult.

To identify web robots' requests, the data cleaning module implements two different techniques.

- In the first technique, all records containing the name "robots.txt" in the requested resource name (URL) are identified and straightly removed.
- The next technique is based on the fact that the crawlers retrieve pages in an automatic and exhaustive manner, so they are distinguished by a very high browsing speed. Therefore, for each different IP address, the browsing speed is calculated and all requests with this value more than a threshold are regarded as made by robots and are consequently removed. The value of the threshold is set up by analyzing the browser behavior arising from the considered log files.

This helps in accurate detection of user interested patterns by providing only the relevant web logs. Only the patterns that are much interested by the user will be resulted in the final phase of identification if this cleaning process is performed before start identifying the user interested patterns.

Computing the reference length: Reference Length is nothing but the time taken by the user to view a particular page. This plays an important role in the following procedures. Generally it is calculated by the difference between access time of a record and the next record. But this is not correct since the time includes data transfer rate over internet, launching time to play audio or video files on the web page and so on. The user's real browsing time is very difficult to analyze. The data transfer rate and size of page is also considered and the reference length is calculated as:

$$RL_{time} = RLT' - bytes\_sent/c$$

Where:

RLT'	=	The difference of access time between a record and the next one
bytes_sent	=	Taken from log entry of a record
c	=	The data transfer rate

User identification: The log file after cleaning is considered as Web Usage Log Set WULS = {UIP, Date, Method, URI, Version, Status, Bytes, ReferrerURL, BrowserOS }.

The next important and complex step is unique user identification. The complexity is due to the local cache and proxy servers. To overcome this cookies are used. But users may disable cookies. Another solution is to collect registration data from users. But users neglect to give their information due to privacy concerns. So majority of records does not contain any information in the user-id and authentication fields. The fields which are useful to find unique users and sessions are:

- IP address
- User agent
- Referrer URL

Users and sessions are identified by using these fields as follows. If two records has same IP address

check for browser information. If user agent value is same for both records then they are identified as from same user.

Session identification: The goal of session identification is to divide the page accesses of each user into individual sessions. These sessions are used as data vectors in various classification, prediction, clustering into groups and other tasks. If URL in the referrer URL field in current record is not accessed previously or if referrer url field is empty then it is considered as a new user session. Reconstruction of accurate user sessions from server access logs is a challenge task and time oriented heuristics with a time limit of 30 min is followed.

From WULS, the set of user sessions are extracted as referrer based method and time oriented heuristics:

$$USS=\{USID,(URI_1,ReferrerURI_1,Date_1)\ldots..(URI_k, ReferrerURI_k,Date_k)\}$$

where, $1 \leq k \leq n$ and n denotes the amount of records in WULS. Every record in WULS must belong to a session and every record in WULS can belong to one user session only. After grouping the records into sessions the path completion step follows.

Path completion: Path completion step is carried out to identify missing pages due to cache and 'Back'. Path Set is the incomplete accessed pages in a user session. It is extracted from every user session set.

Path Combination and Completion: Path Set (PS) is access path of every USID identified from USS. It is defined as:

$$PS = \{USID,(URI_1,Date_1, RLength_1),\ldots (URI_k, Date_k, RLength_k)\}$$

where, Rlength is computed for every record in data cleaning stage. After identifying path for each USID path combination is done if two consecutive pages are same. In the user session if any of the URL specified in the Referrer URL is not equal to the URL in the previous record then that URL in the Referrer Url field of current record is inserted into this session and thus path completion is obtained. The next step is to determine the reference length of new appended pages during path completion and modify the reference length of adjacent ones. Since the assumed pages are normally considered as auxiliary pages the length is determined by the average reference length of auxiliary pages. The reference length of adjacent pages is also adjusted.

Transactions identification: The goal of transactions identification is to create meaningful clusters of references for each user. Transaction identification is done by merges or divides approaches. To find out the user's travel pattern and user's interests, two kinds of transactions are defined. i.e., travel path transactions and content only transactions. The travel path is a combination of auxiliary and content pages accessed by a user. The content only transactions are only content pages which are used in mining to discover user's interest and cluster users visiting the same web site.

There are three methods available to identify transactions; they are identification by Reference Length, identification by Maximal Forward Reference and identification by Time Window.

In the proposed method a combination of all methods are used and Content Path Set and Travel-path transactions are identified. First by using Maximal Forward Reference the paths in a session is split into forward reference paths. Travel paths of a user session are found. Travel Path Set is defined as the set of user travel paths, the member of TPS includes travel paths, the member of TPS includes travel paths having same USID, defined as:

$$TPS = <USID_i, TP_i^1, TP_i^2, \ldots, TP_i^n>$$

where, TP is the travel path is a group of URIs which are arranged according to the access time, a travel path including k URIs is defined as:

$$TP = \{URI_1, URI_2,..., URI_k\}$$

Reference Length algorithm is used to distinguish content pages from auxiliary pages. The algorithm depends on the time spent on viewing a page. A page is identified as content page if it exceeds a cut-off time or as auxiliary page if it is less than cutoff time. Cutoff time is calculated using a formula:

$$t = - \lambda . \ln r$$

where, r is the percentage of content pages in the log found from the site . Normally the last page in every travel path is identified as content pages and leading pages are auxiliary pages. λ is the mean reference length of all pages in the log. In this the last record is ignored since last pages are normally considered as content pages. But it may be auxiliary pages also. To solve this issue the third algorithm Transactions by Time Window is used. In this a default time is fixed for each session and divided the path into transactions. The time difference between the first and last page access is calculated. That is considered as total time of transaction. Then the difference between Time Window and calculated total time is calculated. If the difference is less than cut off time it is considered as auxiliary page or as content page.

From the above techniques content transactions are identified. Content Path Set (CPS) is the set of content pages, used for mining, corresponding to each user session, is written as:

$CPS = < USID_i, CP_i^1, CP_i^2 \ldots CP_i^k >$

where, k is the number of content pages for the ith user session.

RESULTS

The experiments are conducted in the proposed technique by using the log obtained from the reputed college web site for about 30 days in 2010. The obtained record consists of 1000 records in the log file. Then the data cleaning process is carries out (Huiying and Wei, 2004). Initially, after removing records with graphics and videos format such gif, JPEG, 520 records are obtained. Then by checking the status code, the total of 450 records is resulted. Finally, 390 records are resulted after applying robot cleaning process. In the proposed method the records accessed by robots, agents are also cleaned by considering the access time limit of 2 sec. The sample of 5 records are considered and experimented.

Figure 1 shows the time required for determining user interested pattern after different data cleaning techniques. In the sample 1, the total of 1000 records are obtained initially. Then after removing the gif status, 520 records are resulted. Finally 350 records are obtained after robots cleaning. In sample 2, initial record is 950-480 records are resulted after gif status removal and finally 320 records are obtained after robots cleaning process. When considering sample 4, the initial record is 800- 350 records are resulted after gif status removal and finally 250 records are obtained after robots cleaning process. As the number of irrelevant records is discarded, this helps in determining the user interested pattern more accurately in less time.

For sample 1, the time required for prediction using initial log is 119 sec, whereas, 77 sec after cleaning by gif status removal and it takes only 52 sec. For sample 2, only 30 sec is required for determining the user pattern by including robots cleaning and more time is required when the robots cleaning is not included. For sample 3, 106 and 81 sec are required by using original log and log after gif status removed, whereas, only 56 sec is required by using the log after robots cleaning.

After data cleaning, 6 users are identified according to IP addresses, browsers and operating systems. Furthermore, by using the referer-based and the time-oriented heuristics methods, 60 user sessions are distinguished in this experiment. Then the path completion technique is applied in order to determine the path accessed by the user. The path completed for a user by using original log is given in Table 1.

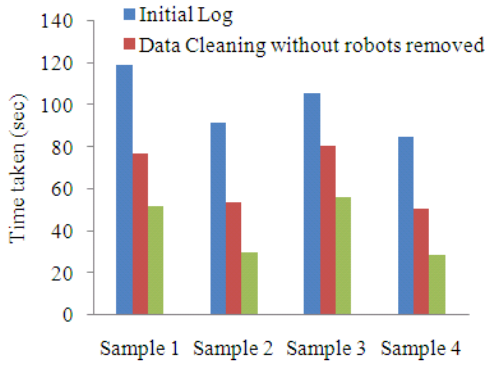

Fig. 1: Time taken for user interested pattern prediction after different data cleaning techniques

Table 1: Path completed for a user by using original log

IP address	User id	Session id	Path completed
116.128.56.89	1	1	16-17-18-17-18-19-20-15-11-25-26-45-22-41
116.128.56.89	1	2	25-26-30-35-41-45-32-11-22

Table 2: Path completed for a user by using log after cleaning but without robots cleaning

IP address	User id	Session id	Path completed
116.128.56.89	1	1	16-17-18-17-18-19-20-15-11-22
116.128.56.89	1	2	25-26-30-35-32-22

Table 3: Path completed for a user by using log after robots cleaning

IP address	User id	Session id	Path completed
116.128.56.89	1	1	16-17-18-17-18-19-20
116.128.56.89	1	2	25-26-30-35

Table 2 shows the path completed (Li *et al*., 2008) for a user by using log after cleaning but without robots cleaning. It can be observed from Table 2 that the irrelevant pages found in Table 1 are eliminated. Finally, Table 3 provides path completed for a user by using log after robots cleaning. From Table 3, it can be observed that only most relevant web pages interested by the user is obtained, whereas, in Table 1-2 some of the irrelevant wed pages are considered for predicting the user interested patterns.

DISCUSSION

The problem in web log mining is solved in this study. Initially, the logs are collected and the preprocessing steps are carried out. The preprocessing steps carried out here are removal of records of graphics, videos and the format information, the records

with the failed HTTP status code, Method field and Robots cleaning. This will help in reduction of quantity of data to be passed to further processing. Then the users are identified by user identification phase. From this the sessions are identified. Next, the path completion step is carried out to identify missing pages due to cache and 'Back'. Path Set is the incomplete accessed pages in a user session. It is extracted from every user session set. Then, the user transactions are identified. Finally, from the obtained data, content path set are identified which will help in better web prediction. Then the experiment is conducted using the log obtained from the reputed college web site to evaluate the proposed technique. The experimental result shows the improvement in the web log mining.

CONCLUSION

A data preprocessing treatment system for web usage mining has been analyzed and implemented for log data. It has undergone various steps such as data cleaning, user identification, session identification, path completion and transaction identification. Data cleaning phase includes the removal of records of graphics, videos and the format information, the records with the failed HTTP status code and finally robots cleaning. Different from other implementations records are cleaned effectively by removing robot entries. The reference length is computed by considering the byte transfer rate. Apart from using Maximal Forward Reference (MFR) and Reference Length (RL) algorithm Time Window concept is also combined to find content pages. Travel path transactions are constructed to know the navigational behavior of users. Content page set is used for analyzing users and so that modification of sites can be done. This preprocessing step is used to give a reliable input for data mining tasks. Accurate input can be found if the byte rate of each and every record is found. The data cleaning phase implemented in this study will helps in determining only the relevant logs that the user is interested in.

REFERENCES

Aghabozorgi, S.R. and T.Y. Wah, 2009. Using incremental fuzzy clustering to web usage mining. Proceedings of the International Conference of Soft Computing and Pattern Recognition, Dec. 4-7, IEEE Xplore, Malacca, pp: 653-658. DOI: 10.1109/SoCPaR.2009.128

Aziz, A.A., M.Y.M. Saman and M.P. Hamzah, 2011. Using metadata analysis and base analysis techniques in data qualities framework for data warehouses. Am. J. Econ. Bus. Admin., 3: 112-119. DOI: 10.3844/ajebasp.2011.112.119

Baraglia, R. and P. Palmerini, 2002. SUGGEST: A web usage mining system. Proceedings of the International Conference on Information Technology: Coding and Computing, Apr. 8-10, Las Vegas, Nevada, pp: 282-282.

Chang-bin, J., 2010. Application of cloud model in personalized service recommendation of web log mining. Proceedings of the International Conference on Biomedical Engineering and Computer Science, Apr. 23-25, IEEE Xplore, Wuhan, pp: 1-4. DOI: 10.1109/ICBECS.2010.5462359

Chen, J., J. Yin, A.K.H. Tung and B. Liu, 2004. Discovering web usage patterns by mining cross-transaction association rules. Proceedings of International Conference on Machine Learning and Cybernetics, Aug. 26-29, Zhongshan University, Guangzhou, China, pp: 2655-2660. DOI: 10.1109/ICMLC.2004.1378232

Dong, D., 2009. Exploration on web usage mining and its application. Proceedings of the International Workshop on Intelligent Systems and Application, May 23-24, IEEE Xplore, Wuhan, pp: 1-4. DOI: 10.1109/IWISA.2009.5072860

Etminani, K., A.R. Delui, N.R. Yanehsari and M. Rouhani, 2009. Web usage mining: Discovery of the users' navigational patterns using SOM. Proceedings of the 1st International Conference on Networked Digital Technology, July 28-31, IEEE Xplore, Ostrava, pp: 224-249. DOI: 10.1109/NDT.2009.5272158

Huiying, Z. and L. Wei, 2004. An intelligent algorithm of data pre-processing in Web usage mining. Proceedings of 5th World Congress on Intelligent Control and Automation, June 15-19, Tianjin University, China, pp: 3119- 3123. DOI: 10.1109/WCICA.2004.1343095

Hussain, T., S. Asghar and N. Masood, 2010. Web usage mining: A survey on preprocessing of web log file. Proceedings of the International Conference on Information and Emerging Technologies, June 14-16, IEEE Xplore, Karachi, pp: 1-6. DOI: 10.1109/ICIET.2010.5625730

Inbarani, H.H., K. Thangavel and A. Pethalakshmi, 2007. Rough set based feature selection for web usage mining. Proceedings of the International Conference on Conference on Computational Intelligence and Multimedia Applications, (ICCIMA'07), IEEE Computer Society Washington, DC, USA., pp: 33-38. DOI: 10.1109/ICCIMA.2007.373

Jalali, M., N. Mustapha, N.B. Sulaiman and A. Mamat, 2008. A web usage mining approach based on LCS algorithm in online predicting recommendation systems. Proceedings of 12th International Conference Information Visualisation, (IV'08), IEEE Computer Society Washington, DC, USA., pp: 302-307. DOI: 10.1109/IV.2008.40

Labroche, N., M.J. Lesot and L. Yaffi, 2007. A new web usage mining and visualization tool. Proceedings of the 19th IEEE International Conference on Tools with Artificial Intelligent, Oct. 29-31, Paris, France, pp: 321-328.

Lee, C.H. and Y.H. Fu, 2008. Web usage mining based on clustering of browsing features. Proceedings of the 8th International Conference on Intelligent Systems Design and Application, (ISDA'08), IEEE Computer Society Washington, DC, USA, pp: 281-286. DOI: 10.1109/ISDA.2008.185

Li, Y., B. Feng and Q. Mao, 2008. Research on path completion technique in web usage mining. Proceedings of the International Symposium on Computer Science and Computational Technology, Dec. 20-22, IEEE Xplore, Shanghai, pp: 554-559. DOI: 10.1109/ISCSCT.2008.151

Liu, L. and J. Liu, 2010. Mining web log sequential patterns with layer coded breadth-first linked WAP-tree. Proceedings of the International Conference of Information Science and Management Engineering, Aug. 7-8, IEEE Xplore, Xian, pp: 28-31. DOI: 10.1109/ISME.2010.271

Liu, L. and J. Liu, 2010. Mining web log sequential patterns with layer coded breadth-first linked WAP-tree. Proceedings of the International Conference of Information Science and Management Engineering, Aug. 7-8, IEEE Xplore, Xian, pp. 28-31. DOI: 10.1109/ISME.2010.271

Maratea, A. and A. Petrosino, 2009. An heuristic approach to page recommendation in web usage mining. Proceedings of the 9th International Conference on Intelligent Systems Design and Applications, Nov. 30-2 Dec., IEEE Xplore, Pisa, pp: 1043-1048. DOI: 10.1109/ISDA.2009.252

Nasraoui, O., M. Soliman, E. Saka, A. Badia and R. Germain, 2008. A web usage mining framework for mining evolving user profiles in dynamic web sites. IEEE Trans. Knowl. Data Eng., 20: 202-215. DOI: 10.1109/TKDE.2007.190667

Nina, S.P., M. Rahman, K.I. Bhuiyan and K. Ahmed, 2009. Pattern discovery of web usage mining. Proceedings of the International Conference on Computer Technology Development, Nov. 13-15, IEEE Xplore, Kota Kinabalu, pp: 499-503. DOI: 10.1109/ICCTD.2009.199

Panich, S., 2010. The shortest path with intelligent algorithm. J. Math. Stat., 6: 276-278. DOI: 10.3844/jmssp.2010.276.278

Shinde, S.K. and U.V. Kulkarni, 2008. A new approach for on line recommender system in web usage mining. Proceeding of the International Conference on Advanced Computer Theory and Engineering, Dec. 20-22, IEEE Xplore, Phuket, pp: 973- 977. DOI: 10.1109/ICACTE.2008.72

Wu, C.H., Y.L. Wu, Y.M. Chang and M.H. Hung, 2010. Web usage mining on the sequences of clicking patterns in a grid computing environment. Proceedings of the International Conference on Machine Learning and Cybernetics, July 11-14, IEEE Xplore, Qingdao, pp: 2909-2914. DOI: 10.1109/ICMLC.2010.5580751

Wu, K.L., P.S. Yu and A. Ballman, 1998. SpeedTracer: A web usage mining and analysis tool. IBM Syst. J., 37: 89-105. DOI: 10.1147/sj.371.0089

Yamin, F.M. and T. Ramayah, 2011. The impact of user knowledge on web search satisfaction. Am. J. Econ. Bus. Admin., 3: 139-145. DOI: 10.3844/ajebasp.2011.139.145

Zhang, J., P. Zhao, L. Shang and L. Wang, 2009. Web usage mining based on fuzzy clustering in identifying target group. Proceeding of the International Colloquium on Computing, Communication, Control and Management, Aug. 8-9, IEEE Xplore, Sanya, pp: 209-212. DOI: 10.1109/CCCM.2009.5267789