Sindh Univ. Res. Jour. (Sci. Ser.) Vol. 48 (1) 201-208 (2016)

SI NDH UNIVERSITY RESEARCH JOURNAL (SCIENCE SERIES)

Extraction of Web Navigation Patterns by means of Sequential Pattern Mining

N. A. MAHOTO[++], A. MEMON*, M. MEMON**, M. A. TEEVNO***

Department of Software Engineering, Mehran UET, Jamshoro Sindh, Pakistan

Received 15th August 2015 and Revised 28rd January 2016

Abstract: The world wide web has become a massive source of information as well as means of online business over the last many years. Business organizational websites serve as communication link between customers and business organizations for the e-business. E-business organizations pursuit effective ways of marketing to fulfill their customers' expectations. Thus, e-business organizations need to identify and evaluate the interests of their potential customers for promoting and increasing their business. Intelligent use of web mining techniques helps them to achieve their targeted goals such as assessing priorities and preferences of their customers. This study presents an approach to identify most frequent web navigational patterns from web logs. These patterns, obtained using well-established data mining disciplines: sequential pattern mining and clustering, help in understanding navigational behavior of the web users. The detected patterns may be used for creating dynamic websites, adverting purpose, user profiles, present and future probability, relationship between behavior and website usability.

Keywords: Sequential Pattern Mining, Web Usage Mining, Clustering, Web Navigation

1. INTRODUCTION

The World Wide Web, over the years, has become the most important and myriad source of information. The expert circles expect it to grow towards building up of a primary information resource for the decades to come (Sujatha, 2011).. In today's technological era, Websites are being used as direct communication link between customers and business organizations. Online business is turning traditional ways of business upside down. Meanwhile, business organizations should routinely evaluate whether their online services are working up to the mark for what they are intended for, if converse, then they can make it adaptive by utilizing data usage mining techniques.

Recent years have witnessed growing research in the field of data mining in general, and web usage mining in particular. Web mining is an application of data mining methods and techniques concerning with web content mining, web structure mining and web usage mining. Web content mining is concerned with the contents or the real data the website was designed to convey to users. Specific applications of web content mining include content-based ranking and content-based categorization of web pages (Gupta, and Han, 2012).. Web structure mining analyzes the way the web sites are structured. Link based classification of websites, restructuring of websites and website ranking based on content and structure are the typical applications of web structure mining (Padala, *et al.,* 2013). Web usage mining is the area of web mining that relates to the retrieval of useful information from web log data and assists in the identification of patterns from user's navigational data. Web usage mining can be termed as click stream analysis. Web mining typically examines data sources such as contents of websites (i-e text and graphics etc.), web log data (for example date and time of web navigation, IP addresses), web structure data (i.e., XML and HTML files) (Padala, *et al.,* 2013).

Intelligent use of web usage mining techniques put business corporations in a better position to assess the priorities and preferences of customers and to run the business profitably. E-business organizations are pursuing ways to use web log data for effective marketing and to fulfill customer needs by analyzing customer navigation patterns and buyer's activities, thus consequently increasing production and revenues (Sainath, *et al.,* 2012). The sequential web patterns obtained by analyzing the customer's navigational patterns may offer advice for the e-companies to better improve the website and make it adaptive as it meets customer needs.

This paper focuses on first identifying useful patterns from the active user web logs and then classifies customers according to their interests (in terms of similar web page visits). Several techniques are available to identify patterns from web log data like decision tree, rule based analysis (association rules, sequencing etc.). Decision trees are simple to use and understand but their limitation is these methods do not take time-series data into account. Thus, decision tree technique is unable to determine at which point of time a customer visited certain website. Association rules

[++]Corresponding author. Email: naeemmahoto@gmail.com ; Tel.: +92-333-7538991

* Computer Science, IBA Community College Dadu, Sindh, Pakistan

**Department of Software Engineering, Mehran UET Jamshoro, Sindh, Pakistan

***Department of Electronic Engineering, Mehran UET Jamshoro, Sindh, Pakistan

may provide a useful insight regarding the customer behavior (Sujatha, 2012). For example use of association rules may help a manager to find that the customers buy oil from a supermarket would also purchase vegetables and eventually enable him to arrange supermarket in a better way to attract maximum customers. Sequential patterns can also be used to extract user's navigational behavior. This paper derives navigational patterns of web users from their web logs using sequential pattern mining technique. The web logs have been taken from online data repository. The applied approach also determines the similar groups of web users (users' having similar pattern of navigational patterns). Grouping together similar web users may help e-business organizations to categorize customers according to their behavior and interests.

The remaining paper is organized as follows. Section 2 reports related research; Section 3 thoroughly covers the proposed approach. Section 4 presents experimental results. Finally, Section 5 gives the conclusion and proposed future work.

2 RELATED WORK

Many researchers have devoted their research studies in the field of usage mining of web. The work in (Sujatha, 2012) presented Prediction of web user navigation patterns using Clustering and Classification (PUCC) from web log data. The study (Gnanavel, *et al.,* 2012) about pattern analysis of web usage data suggested for customization i.e. web visualization about how different web usage mining methods can be applied. (Weichbroth *et al.,* 2012), proposed a system framework for mining user navigation behavior and to evaluate the effectiveness of the implemented algorithm focused on critical factors. In 2013, (Singh *et al.,* 2013) carried out experimental work on NASA web server log data to extract interesting and useful knowledge and is analyzed with Web Log Explorer. The focus of (Sharma, *et al.,* 2012). remained on usage patterns of an educational institutions' web access log data for one-month period by identifying number of visitors and to improve web usability.

In 2012 (Dohar *et al.,* 2012) introduced a new session reconstruction algorithm, which produces user navigations as well as the web log data reserved by the web server. In his work, the actual web user sessions will be identified and the successes of different techniques can be evaluated.

An approach for discovering web user's navigation patterns and analyzing of web usage data using Weighted Fuzzy Possibilistic C-Means (WFPCM) for clustering and Adaptive Neuro-Fuzzy Inference System with Subtractive Algorithm (ANFIS-SA) for classification has been introduced in (Vellingiri, *et al.,* 2015). In 2013, (Shanthi *et al.,* 2013) introduced a strategy for automatic discovery of web server log information using web page collection algorithm. Furthermore, the study (Shanthi *et al.,* 2013) for web log analysis also focused on the efficient application of the Web Mining Algorithm. The approach applied in this study uses sequential pattern mining technique to identify the sequence of navigational patterns of web users. The extracted patterns would help e-business organizations to understand the customer interests and behavior in terms of web page visits.

3. EXTRACTING WEB NAVIGATIONAL PATTERNS

The main idea of the proposed research is to create clusters of web users having similar mindset of browsing through the World Wide Web. **(Fig 1)** illustrates the proposed approach, which converts the web log data into subgroups of web users having similar web navigational behavior. It consists of four main blocks. The detailed insight into these blocks and the role of every block is given in the succeeding sections.

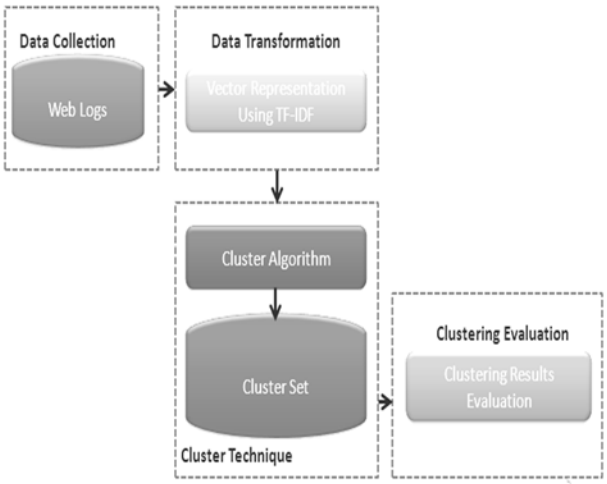

Fig. 1 A proposed approach for web users' navigation pattern determination

3.1 Data collection and preparation

The raw dataset of web users' is obtained from UCI online repository (UCI Machine Learning Repository (2015). The collected dataset in the form of web log files is not suitable for extracting patterns and clustering techniques, therefore, it is processed first and then transformed into a format suitable for clustering by removing unnecessary and noisy data. For example, the missing records are removed and identical records are managed to present distinct records.

3.2 Data Transformation – Vector Representation

This block transforms the web log data into a vector space. The vector space representation is carried out using TF-IDF (Term frequency-Inverse document

frequency) scheme (Pang-Ning, *et al.,* 2006).Each web user is represented through a vector in that vector space model. Every individual web user (i.e., vector) represents the number of occurrences of the navigations (of web pages) visited by the user. It becomes inappropriate to completely characterize a user through an absolute frequency rather the pragmatic way to do this is to represent the user's web log data into weighted exam frequency with TF-IDF scheme. The similar scheme has been used in (Antonelli, *et al.,* 2013) to represent healthcare transactional dataset.

Formally, consider $\boldsymbol{D}$ be the collection of web user records and $\sum n = \{p_1, \dots, p_k\}$ is the set of navigational web pages at least once visited by the web users in $\boldsymbol{D}$. Each web user $user_i$ in $\boldsymbol{D}$ is represented into weighted frequency vector V_{user_i} of $|\sum n|$ web pages. Each element $V_{user_i}[j]$ of vector V_{user_i} expresses weighted frequency $W_{user_i}p_j$ of web page p_j visited by web user $user_i$. For instance:

$$V_{user_i} = \left[W_{user_i}p_1, \dots, W_{user_i}p_{|\sum n|}\right] \qquad \textbf{Eq. 1}$$

The TF-IDF weight $W_{user_i}p_j$ for the pair $(user_i, p_j)$ is the product of $\left(TF_{user_i,p_j}\right)$ and $\left(IDF_{p_j}\right)$.

$$W_{user_i}p_j = \left(TF_{user_i,p_j}\right) * \left(IDF_{p_j}\right) \qquad \textbf{Eq. 2}$$

The Term-Frequency TF_{user_i,p_j} for pair $(user_i, p_j)$ shows relative frequency web page p_j visited by web user $user_i$, which is given by:

$$TF_{user_i,p_j} = {f_{user_i,p_j}}\Big/{\sum_{1\le k\le|\sum|} f_{user_i,p_k}} \qquad \textbf{Eq. 3}$$

where f_{user_i,p_j} is the number of times $user_i$ visited p_j web page and $\sum_{1\le k\le|\sum|} f_{user_i,p_k}$ is the total number of web pages visited by the $user_i$ web user. The Inverse Document Frequency (IDF_{p_j}) for p_j web page expresses the frequency of p_j in web user collection $\boldsymbol{D}$, which is given by:

$$IDF_{p_j} = \log\left[\left||D|/user_i \in D: f_{user_k,p_j} \neq 0\right|\right] \qquad \textbf{Eq. 4}$$

where $|D|$ is the number of web users in web user collection $\boldsymbol{D}$ and $|user_i \in D: f_{user_k,p_j} \neq 0|$ is the number of web users who (at least once) visited p_j web page. Mathematically, the log function does not matter and yields a constant multiplicative factor to produce the overall result.

3.3 Cluster Technique

Web user clustering is important tasks for mining usage data, whose goal is to place the users into "clusters" such that similar users (having same or identical browsing behavior) can be grouped together in one group and dissimilar users are placed in different groups. The proposed approach uses partitioned clustering technique (i.e., k-means (Juang, *et al.,* 1990). which has been used along with other clustering techniques in (Mahoto, *et al.,* 2013).. The k-means (Juang, *et al.,* 1990). is the simplest unsupervised learning algorithm suitable for this study because of the dataset used, which is highly overlapping. The k-means clustering follows a simple, highly efficient and robust procedure with iterative generation of clusters. It actually tries to find out *k* non-overlapping clusters. A cluster centroid, typically the mean of the point in that cluster represents a cluster. The process of cluster generation through k-means starts with arbitrary selection of *k* number of centroids. Then each point from the dataset is assigned to the closely matching centroid on the basis of an appropriate proximity function selected. The centroids for each cluster are continuously updated until the centroids do not change. In a nutshell, k-means performs two steps (1) to take each object belonging to the related group whose mean value similar to that object (2) re-calculate new centroids using mean value. This loop continues until no more changes occur in k-centroids (Xu, *et al.,* 2010),

3.3.1 Distance Measurement

The similarity and dissimilarity between the objects (i.e., web users in this study) requires few measurements. Numerous methods and techniques of similarity measures and distance metrics are available to compute the distance between objects during clustering depending on the nature of their dimensions like Manhattan distance, the Euclidean distance, the hamming distance, Chebyshev distance, and the cosine distance and type of data to perform this requisite task.

When the data points are in the Euclidean space, then it is quite often that Euclidean distance (Eq. 5) metric is used to calculate the distance between objects through the following relation, where p and q are the vectors of m dimensions.

$$d_{E(p,q)} = \sqrt{\sum_{i=1}^{m}(p_i - q_i)^2} \qquad \textbf{Eq. 5}$$

In order to determine the distance from one object to another if the grid-like path is followed, a variation of the above method (i.e., Eq. 5) can be used, and that is termed as Manhattan distance (Eq. 6).

$$d_{M(p,q)} = \sqrt{\sum_{i=1}^{m}|p_i - q_i|} \qquad \textbf{Eq. 6}$$

The Chebyshev distance (Eq. 7), also termed as maximum value distance determines the maximum distance between two objects among all its dimensions. This method is normally appropriate when the distance should reflect the differences in individual dimensions rather than all the dimensions considered together.

$$d_{Ch(p,q)} = \max_{i=1}|p_i - q_i| \quad \textbf{Eq. 7}$$

The Hamming distance (Eq. 8) is useful for categorical values. It also evaluates the number of dimensions at which data points have a different value. The formula for Hamming distance calculation is given below.

$$d_{H(p,q)} = \sum_{i=1}^{m} H(p_i, q_i) \quad \textbf{Eq. 8}$$

The cosine similarity (Eq. 9) computes the cosine of the angle that is formed by two objects regarded as vectors:

$$cosine(p,q) = \frac{\sum_{i=1}^{m} p_i q_i}{\sqrt{\sum_{k=1}^{m} p_i^2}\sqrt{\sum_{k=1}^{m} q_i^2}} \quad \textbf{Eq. 9}$$

In this study, cosine similarity is used because it provides three different measurements such as independent, equal and opposite between two vectors (i.e. vectors of web navigators).

Cosine similarity measurement value lies in the range of -1, 0 and +1. If the cosine measure of vectors is 1, it shows the vectors point in the same direction, if it is equal to -1, it means vectors are in opposite direction and the vectors are independent if the cosine similarity measurement of both the vectors is zero.

3.4 Cluster Evaluation

Indeed it is hard to answer the question that how to judge the quality of clustering results since there is no prior information available. The techniques measuring the quality of clustering outcomes can be categorized as external indexes and internal indexes.

The internal indexes permit the evaluation of clusters when there is no previous information available while for evaluating the cluster sets when solutions are available, external indexes are used. The most commonly employed internal indexes are separation, homogeneity, and silhouette (Eq. 10).

Since no prior information regarding the web users' browsing behavior is available, the proposed approach uses internal indexes (silhouette index (Rousseeuw, 1987). to evaluate the cluster sets. Silhouette index is actually a graphical method for the evaluation and validation of cluster sets. The silhouette index value confirms the correct placement of an object (i.e. web users in this study). Silhouette plot analysis is very much beneficial in the evaluation of distance between resulting clusters. The silhouette plot is actually a measure of the closeness of each point in one cluster to the corresponding point in another cluster, thus providing a method to ascertain parameters like number of clusters. This analysis tool provids numbers that range from -1 to +1 through zero. The negative value signifies the incorrect placement of objects (i.e. web user in this case), while a positive value of silhouette portrays correct web user placement and finally zero shows the objects being placed at border of cluster.

$$Silhouett_i = \frac{b(x)-a(x)}{\max\{a(x),b(x)\}} \quad \textbf{Eq. 10}$$

4. EXTRACTING PATTERNS

There are ample convincing arguments available that frequent pattern mining is efficient only when it mines not all frequent patterns rather only the closed ones because it consequently provides compact and complete result set. BI-Directional Extension (BIDE) (Wang, and Han, 2004). is among the ones being efficient at discovering frequent and similar sequential patterns without candidate maintenance. Proposing a new paradigm in the closed sequential patterns discovery, a new technique called BIDE Extension (forward extension and back word extension) has evolved. The developing of prefix pattern as well as checking its closure, forward extension can be used, while testing both, closure of prefix pattern and pruning the search space, backward extension is needed. Bi-directional extension, by using BackScan pruning method and scan-skip optimization prunes the search space relatively deeper as compared to previously developed algorithms. In a nutshell, BIDE surpasses all the previous algorithms in various areas like it runs a bit faster, consumes considerably less memory and especially when the support value is low in magnitude, it is linearly scalable in terms of database size (Wang, and Han, 2004). This study exploits BIDE algorithm to extract navigation patterns. The BIDE algorithm implementation code used for generating sequential web user patterns has been kindly provided by Philippe Fournier-Viger (Fournier- *et al.*, 2014).

5. RESULTS AND DISCUSSION

This section attempts to summarize the results obtained while analyzing and applying the proposed approach on the web log data taken from UCI Machine Learning Repository (UCI Machine Learning Repository (2015). The implementation of clustering process and cluster evaluation is carried out through RapidMiner tool (Rapid Miner Project, 2015). while for pre-processing steps python programming language (Python Software Foundation, (2015) is employed. RapidMiner, as well as python programs both can run on several computer platforms including Mac OSX, Linux and Windows.

5.1 Web Navigation Patterns before clustering

The sequential patterns, obtained by applying the BIDE algorithm a well-known sequential pattern mining algorithm, are shown in **(Fig 2).** These patterns indicate

that, for example, 79% of the web users visited front-page while 65% of the users visited news web-page.

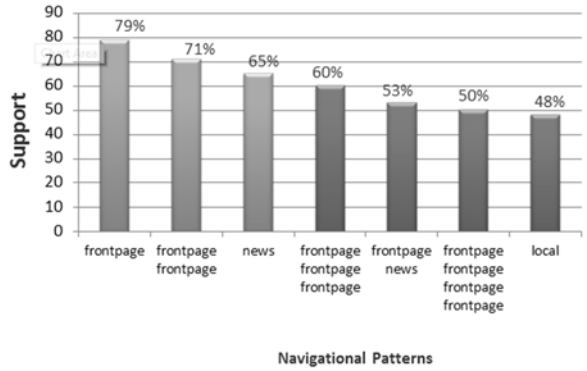

Fig.2 Navigation Patterns before clustering

These sequential patterns do not result in the formation of groups or clusters of the users having similar behavior of navigation or surfing through internet. Therefore, in order to make clusters of web users so as to divide them in different groups in which users having similar navigational behavior fall in one specific group. A detailed overview of the clustering techniques and the method used this approach was highlighted in section 3.3. The following **(Table 1)** presents silhouette values of users in different clusters after applying clustering technique.

Table: 1 k-means clustering and their quality index

	C0	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11
Users'	4120	2393	4185	7614	3618	9860						
Silhouette	0.32	0.47	0.41	0.22	0.51	0.12						
Users'	3294	3911	2934	4339	2239	4834	1558	8681				
Silhouette	0.52	0.41	0.50	0.25	0.45	0.35	0.60	0.11				
Users'	2082	4349	2794	1534	3093	5452	3433	3588	1998	3468		
Silhouette	0.46	0.35	0.51	0.61	0.52	0.04	0.37	0.31	0.36	0.41		
Users'	4099	2005	2445	3034	1870	3129	2912	2758	3008	2045	1525	2960
Silhouette	0.34	0.46	0.30	0.42	0.36	0.18	0.35	0.50	0.37	-0.23	0.61	0.52

Silhouette values range is [-1 +1]. If the silhouette value is high, it signifies that the web users are closely matched to their own cluster and vice versa.

5.2 Web Navigation Patterns after clustering

In the next phase, the cluster sets obtained from the clustering technique are put to evaluation, the process called as validation. The cluster sets subsequently are evaluated by means of quality index (see Section 3.4). The homogeneity and heterogeneity between clusters is balanced by silhouette quality index. The final outcome of the validation is actually the best clustering result of the silhouette index.

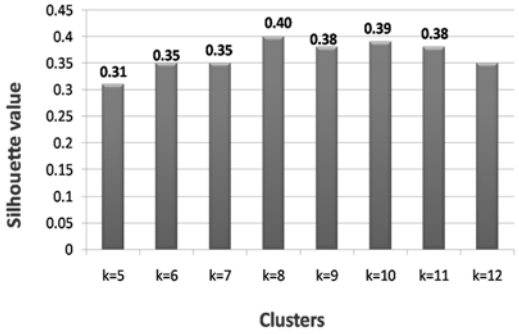

Fig. 3 Evaluation of clusters using Silhouette quality index

In **(Fig 3)**, it can be clearly observed that there are different silhouette index values at different values of k. It is evident from the figure that k-means clustering algorithm produces far superior results when k=8, put it in another way when the number of clusters equals to 8 the k-means clustering algorithm produced desirable and superior results in terms of silhouette quality index. Besides that the results obtained from others values of k (i.e. k=6, k=8, k=10, k=12) are not taken into consideration because their silhouette quality indexes are low.

The following table provides the whole crux of the mined navigational patterns of the considered dataset. This **(Table 2)** highlights frequency of web pages surfed through by web users in each cluster and the bold values distinctly indicate the highest and most frequently accessed web page in any particular cluster.

Table: 2 Clustering and Navigational pattern results

Web pages	C0	C1	C2	C3	C4	C5	C6	C7
Front page	65%	95%	60%	78%	62%	82%	42%	**95%**
msn-sports	-	-	-	-	-	-	**100%**	-
misc	49%	25%	59%	65%	54%	**90%**	56%	19%
weather	16%	-	19%	-	**100%**	-	27%	-
On-air	70%	24%	46%	**95%**	40%	31%	34%	27%
msn-news	-	-	**100%**	-	23%	-	50%	-
sports	28%	**100%**	20%	-	25%	16%	75%	21%
Summary	**100%**	-	-	16%	-	-	-	-

For instance in cluster 1 (i.e., C1), the maximum accessed webpage by users is sports related news, while in cluster 4 (C4), web users are most interested in knowing about weather conditions.

If the individual clusters are examined, for instance, it is observed that maximum (100%) of the users browsed through summary news, 85% of the users accessed summary then again clicked on summary, while 65% of users accessed front page in C0. In this cluster (i.e. C0) no one was interested in msn-sports or msn-news.

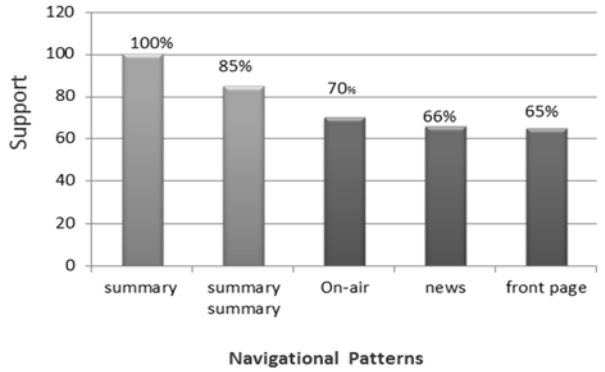

Fig. 4 Sequential patterns of C0

In C3, maximum number of web users (95%) seemed to be interested in on-air web page, 71% visited front page and then accessed on-air (i.e., <front page, on-air>), whereas 65% accessed misc, 78% front page and no one in this cluster visited weather or sports.

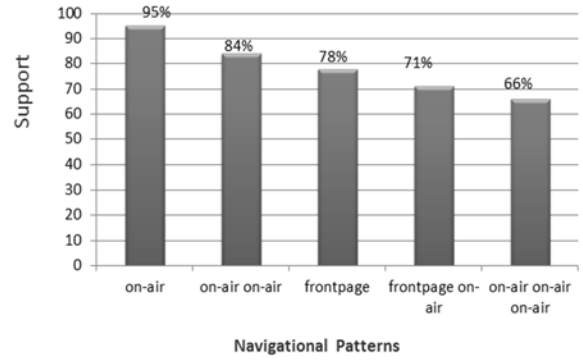

Fig.5 Sequential patterns of C3

6 CONCLUSION AND FUTURE WORK

In recent years, web mining has become an indispensable tool especially in the field of electronic commerce. With the tremendous advancement of web-based applications, issues related to online business intelligence such as domain knowledge of e-commerce, evolution of e-commerce (particularly in online business) necessitates the use of web mining techniques and strategies to cope up with the dynamic challenges to e-business. Web access log data is abundantly available from the World Wide Web, which needs to be transformed into some meaningful format, for instance, obtaining web users' preferences and their navigational behavior to understand the in depth preferences and interests of web users.

In this study, web usage mining has been utilized to predict users' web navigational behavior. Specifically, two main techniques, cluster analysis and sequential pattern mining, have been used to get the desired navigational patterns from real web log datasets of MSNBC (UCI Machine Learning Repository (2015).

The experimental results clearly report that the proposed approach in determining the users browsing behavior is quite pragmatic, effective, feasible and having scalability. The obtained navigational patterns may be used for several purposes such as personalization, advertising, proper website maintenance, prediction of marketing trends, product delivery, enhancing the competitive strength of enterprises, enhancing marketing strategy and getting market data.

REFERENCES:

Antonelli, D., E. Baralis, G., Bruno, T. Cerquitelli, S. Chiusano, and N. A. Mahoto, (2013). Analysis of diabetic patients through their examination history. Expert Systems with Applications. Vol. 40(11), 4672–4678

Dohare, M. P. S., P. Arya, and A. Bajpai, (2012). Novel Web Usage Mining for Web Mining Techniques. International Journal of Emerging Technology and Advanced Engineering, 2(1), 253-262 .

Fournier-V P., A. Gomariz, T. Gueniche, A. Soltani, C. Wu., V. S. Tseng, (2014). SPMF: a Java Open-Source Pattern Mining Library. Journal of Machine Learning Research (JMLR), 15: 3389-3393

Gupta, M., and J. Han, (2012). 'Applications of pattern discovery using sequential data mining'. In Pattern Discovery Using Sequence Data Mining: Applications and Studies / Pradeep Kumar, P. Radha Krishna, S. Bapi Raju. Information Science Reference (IGI Global). 1-23. (it's Book Chapter)

Gnanavel, M., and E. R Naganathan, (2012). A study of pattern analysis techniques of web usage. International Journal of Web Technology, 1(1), 5-10.

Juang, B. H., and L. R. Rabiner, (1990). The segmental K-means algorithm for estimating parameters of hidden Markov models. Acoustics, Speech and Signal Processing, IEEE Transactions on, 38(9), 1639-1641.

Mahoto, N. A., F. K. Shaikh, and A. Q Ansari, (2013). Exploitation of Clustering Techniques in Transactional Healthcare Data. Mehran University Research Journal of Engineering and Technology, 33(1), 77-92.

Padala, V. K., S. Yasin, and D. B Alanka, (2013). A Novel Method for Data Cleaning and User-Session Identification for Web Mining. International Journal of Modern Engineering Research (IJMER), 3(5), 2816-2819.

Python Software Foundation, P. S. (2015). Python Programming Language Official Website. URL: http://www.python.org/ Last access on August 2015

Pang-Ning, T., M., Steinbach, and V. Kumar, (2006). Introduction to data mining (2nd Edition). Addison-Wesley Longman Publishing Co. (it's Book, Publisher: Addison-Wesley Longman Publishing Co.)

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics, 20, 53-65.

Sainath, N., V. Narayandas, U., Moulali, and B Rao, (2012). K. Deployment of Novel Techniques on Web Log Data to Analyze User Navigation Patterns. International Journal of Engineering and Innovative Technology (IJEIT), 1(4), 56-60.

Sujatha, V. (2012). Improved user Navigation Pattern Prediction Technique from Web Log Data. Procedia Engineering, 30, 92-99.

Sujatha, M. V (2011).A Study of Web Navigation Pattern Using Clustering Algorithm in Web Log Files. International Journal of Scientific and Engineering Research, 2(9), 1-5.

Singh, N., A. Jain, and R. S. Raw, (2013). Comparison Analysis Of Web Usage Mining Using Pattern Recognition Techniques. International Journal of Data Mining and Knowledge Management Process (IJDKP), 3(4), 137-147.

Sharma, A. K., and P. C. Gupta, (2012). Identifying the number of visitors to improve website usability from educational institution web log data. International Journal of Computer Applications Technology and Research, 2(1), 22-26.

Shanthi, R., and S. P. Rajagopalan, (2013). An Efficient Web Mining Algorithm To Mine Web Log Information. International Journal of Innovative Research in Computer and Communication Engineering, 1(7), 1491-1500.

UCI Machine Learning Repository (2015). URL: https://archive.ics.uci.edu/ml/datasets.html Last Accessed on August 2015

Vellingiri, J., S. Kaliraj, S., Satheeshkumar, and T. Parthiban, (2015). A Novel Approach for User Navigation Pattern Discovery and Analysis For Web Usage Mining. Journal of Computer Science, 11(2), 372-382.

Weichbroth, P., M. Owoc, and M.Pleszkun, (2012), September). Web user navigation patterns discovery from WWW server log files. In proceeding of IEEE Federated Conference on Computer Science and Information Systems, IEEE, Poland, pp-1171-1176. (Location: Poland, Publisher: IEEE)

Wang, J., and J. Han, (2004). BIDE: Efficient mining of frequent closed sequences. In Data Engineering, 2004. Proceedings. 20th International Conference on Data Engineering, IEEE, Boston, MA, USA, pp-79-90. (Location: Boston, MA, USA, Publisher: IEEE)

Xu, J., and H. Liu, (2010). Web user clustering analysis based on KMeans alorithm. International Conference on Information, Networking and Automation (ICINA), vol. 2, 6-9.