IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Issue-2, Feb 2019 E-ISSN: 2347-2693

Frequent Sequential Pattern Mining in Web Log Data – A Simple Approach

A. Saravanan[1], S. Sathya Bama[2]*

[1]Department of MCA, Sree Saraswathi Thyagaraja College, Tamil Nadu, India – 642205
[2] Independent Researcher, Coimbatore, India

*Corresponding Author: ssathya21@gmail.com

DOI: https://doi.org/10.26438/ijcse/v7i2.2126 | Available online at: www.ijcseonline.org

Accepted: 12/Feb/2019, Published: 28/Feb/2019

***Abstract*—** With nurturing reputation of the World Wide Web, large quantity of web usage data is collected by the web servers and stored in web access log files. Web usage mining is a technology to mine valuable knowledge from the World Wide Web. It intends to discover interesting user access patterns from web log files. Analysis of these user access patterns is used to determine that the information architecture of the web site can be reorganized to better facilitate information retrieval. Association rule mining is also used to find association relationships amongst large data sets. Mining frequent sequential patterns is a significant aspect in association rule mining. Based on this, the changes can be suggested to the web site by placing embedded hyperlinks on the home page to the frequently accessed sections of the web site. In this paper, a very efficient algorithm has been adopted for expert systems to mine frequent sequential patterns in web usage data.

***Keywords*—** Data Mining, Web Mining, Association rule mining, Frequent sequential pattern mining, Web Log files, Web usage mining

I. INTRODUCTION

Due to massive growth of the World Wide Web, the enormous amount of data is now freely available to user access. Managing and organizing these data is a necessary task to access the data efficiently. To increase the performance of web sites, the changes should be made as per users' interests. Though, numerous data mining techniques which are used to learn the hidden information from the web are available, they do not discover the entire knowledge from the web. For efficiency, along with the data mining techniques, other methods such as artificial intelligence, information retrieval and natural language processing techniques can be combined for the betterment of web mining [1]. According to analysis, web mining is broadly classified into three different categories [2, 3, 4, 5], which are Web Usage Mining, Web Content Mining and Web Structure Mining. Figure 1 portrays the categories of web mining.

Web usage mining is the practice of extracting useful information from server logs and finding out what users are looking for on the Internet. Web structure mining is the process of using graph theory to study the connection structure of a web site. Web content mining is mining, extraction and integration of useful data, information and knowledge from Web page contents. Thus, Web mining has been developed into an independent research area. The focus of this paper is to provide a modern approach for frequent sequential pattern mining for discovering different types of patterns in a Web log database.

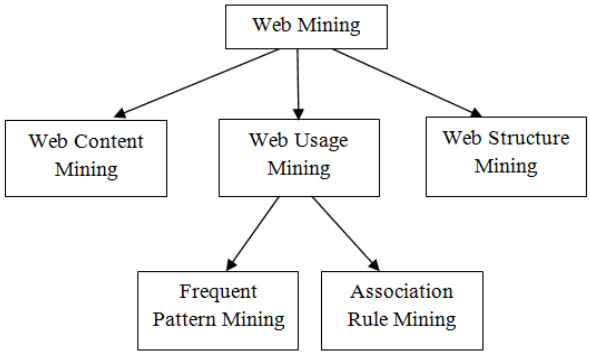

Figure 1. Web Mining Categories

The focus of this paper is to provide a modern approach for frequent sequential pattern mining for discovering different types of patterns in a Web log database.

The organization of the paper is as follows: The importance of mining of web log files is explained in section II. Section III presents the review of literature. Proposed methodology is

described in section IV. Section V presents the experimental results. Section VI describes the conclusion and future enhancement.

II. MINING THE WEB LOG FILES

Web usage mining is the task of identifying the behaviors of the users while they are searching and navigating the Web. The goal of understanding the navigation preferences of the web site visitors is to personalize the Web portals [6] or to improve the web site structure and performance of the Web server [7]. The information offered by the server or client log files can be used to create several concepts like users, page-views, click-streams, and server sessions [8, 9]. A user is defined as an individual who access the files from a web server through a web browser. A page-view consists of several items like frames, text, graphics and scripts with a single user action like mouse click in a web browser to construct a single web page [9].

A click-stream is a sequential series of page-view requests made by a user during navigation. A user session is defined as the time-delimited set of page-views across the entire Web [10]. A server session is defined as the subset of the user session for a specific web server [11]. The process of Web usage mining can be divided into three phases: preprocessing, pattern discovery, and pattern analysis [1, 8]. These phases are depicted in the Figure 2. The first phase is the preprocessing phase which consists of converting usage information from various log files into the data abstraction necessary for pattern discovery [8]. The data recorded in server logs reveal the concurrent access of a Web site by multiple users. The Web server can also store other kinds of usage information such as cookies, which are generated by the Web server for individual client browsers to track the site visitors [12, 13].

After identifying the users, the click-stream for each user must be divided into sessions. Hence, the output of the conversion in preprocessing phase can be used as the input for the algorithms. The next phase in web usage mining is the pattern discovery phase. New methods and algorithms have been used in this phase. This phase has two main operations: association (i.e. which pages tend to be accessed together), and sequential analysis (the order in which web pages tend to be accessed) [12].

Frequent patterns are subsequences of itemsets that appear in a dataset with frequency not less than a user-specified threshold. For example, a set of items, such as milk and bread, which appear frequently together in a transaction dataset, is a frequent itemset. A subsequence, such as buying first a PC, then a digital camera, and then a memory card, if it occurs frequently in a database, is a (frequent) sequential pattern.

Finding these frequent sequential patterns plays an essential role in mining associations, correlations which needs the Web pages that are visited by a given user. To find frequent sequential pattern, the original ordering of the pages is also important, and if a page was visited more than once by a given user in a user defined time interval, then it is relevant as well [1].

The last phase is the pattern analysis phase. This phase mainly concentrates on discarding the uninteresting knowledge or patterns found in the previous phase and reveling conclusions. Visualization techniques are useful to help in analyzing the discovered patterns.

This paper mainly focuses on the frequent sequential patterns that are visited by the users and mining association rule from the sequential pattern. Association rules in Web logs are discovered in [14, 15, 16, 17, 18].

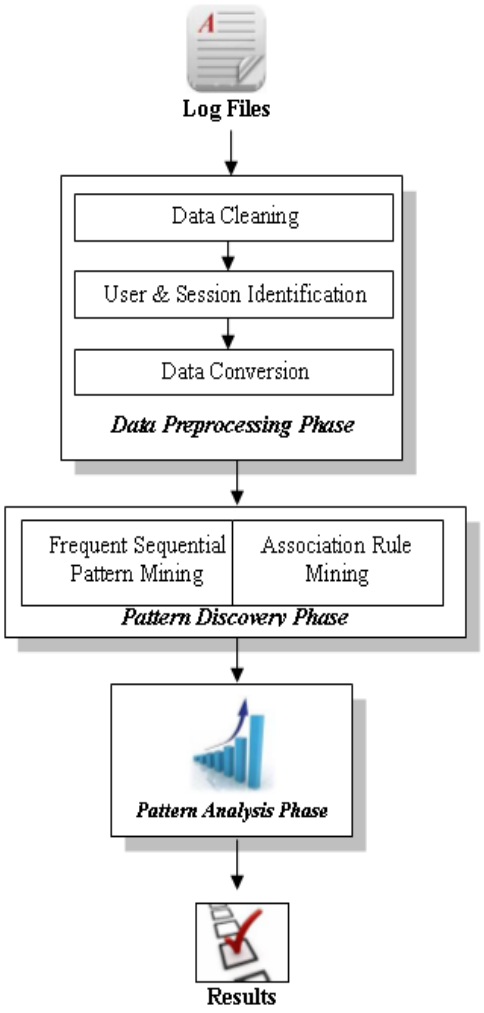

Figure 1. Web Usage Mining Process

III. RELATED WORK

A) Association Rule Mining

Association rules are used in order to discover the pages which are visited together even if they are not directly connected, which can disclose associations between groups of users with specific interest [6]. This information can be used for example for restructuring Web sites by adding links between those pages which are visited together. Association rules in Web logs are discovered in [16, 17, 18,7].

B) Frequent Pattern Mining

Most of the algorithms are based on classical algorithm of association rule mining [19, 20, 21]. Lots of algorithms for mining association rules and their alteration are proposed on the basis of Apriori Algorithm [22]. Most of the previous studies adopt Apriori-like algorithms, which generate-and-test candidates to find frequent patterns. Recently, different works have been proposed a new way to mine patterns in databases [19].

FP-growth method mines the complete set of frequent itemsets without candidate generation [23]. Both the Apriori and FP-growth methods mine frequent patterns from a set of transactions by scanning the horizontal database format again and again. PrefixSpan works in a divide-and-conquer way. The first scan of the database derives the set of length-1 sequential patterns. Its general idea is to examine only the prefix subsequences and project only their corresponding postfix subsequences into projected databases. In each projected database, sequential patterns are grown by exploring only local frequent patterns [24, 25]. But, none of the algorithms filter or reduce the database in each pass of apriori algorithm to count the support of prune pattern candidate from database. We propose a new novel algorithm for frequent pattern mining that uses database as a matrix representation which reduces the number scans in a database [26, 27].

C) Sequential Pattern Mining

The sequential pattern mining problem was first introduced by Agrawal and Srikant in [20]: Given a set of sequences, where each sequence consists of a list of elements and each element consists of a set of items, and given a user-specified min support threshold, sequential pattern mining is to find all of the frequent subsequences, i.e., the subsequences whose occurrence frequency in the set of sequences are not less than the min support. Sequence mining can be used for discovering the Web pages which are accessed immediately after another. Using this knowledge, the trends of the activity of the users can be determined and predictions to the next visited pages can be calculated. Sequence mining is accomplished in [28, 29, 30].

IV. PROPOSED ALGORITHM

Frequent Sequential Itemset Mining is used to discover useful patterns in customers' transaction databases. A customers' transaction database consists of sequence of transactions (T = t1. . . tn), where each transaction is an itemset. An itemset with k elements is called a k-itemset. The support of an itemset X, denoted as sup(X), is the number of those transactions that contain X. An itemset is frequent if its support is greater than a support threshold given by the user and is denoted by min_sup. Next to the support and confidence measures, a lot of other interestingness measures have been proposed in order to get better or more interesting association rules.

The frequent sequential itemset mining problem is to find all frequent itemset in a given transaction database in a sequential manner. The task of discovering all frequent itemsets is quite challenging. The search space is exponential in the number of items occurring in the database. The support threshold limits the output to a hopefully reasonable subspace. Also, such databases could be massive, containing millions of transactions, making support counting a tough problem.

This mining algorithm mines the entire database and finds the sequential frequent patterns. At each level the database is projected as a table consists of item id, transaction id string and the count, the number of occurrence of the item in the database. First the algorithm finds out the frequent-1 items C_1 with transaction id string and count. In the next step construct the F-2 matrix by scanning the database once. The values of the matrix are the transaction id string. For each level, generate the candidate pattern C_k, by making use of the matrix and frequent-1 itemset. For each candidate pattern C_k find out the transaction id string from the F-2 matrix instead of scanning the entire database. For example, the transaction id string for I2I3I4 will be I2I3 ∩ I3I4. Prune the candidate pattern C_k, and find out the frequent k itemset L_k. Use this L_k itemset to generate the candidate itemset for the next level. This will make the process faster. Continue this process until the frequent itemset or if at all not possible to generate the candidate itemset for the next level.

From these frequent sequential patterns, association rules can be mined. The database with sample records consists of transaction id as index and the itemset string as shown below.

Table 1. Datatbase D

Transaction ID	Sequence of web pages
T1	3,2,1,3
T2	2,1,3
T3	3,2

FSPM Algorithm

Step 1: Scan the database once and identify the maximum length of the transaction δ

Step 2: Compute Sequential Frequent-1 Items (L_1) from the database

Step 3: Construct the F-2 matrix with the frequent-1 items from the original database.

Step 4: Compute Sequential Frequent-2 Items(L_2) from the matrix

Step 5: K =3

While K< δ

While $L_{k-1} \neq \{\}$ do

Compute C_k of all candidate k-1 itemsets using the F-2 matrix and L_1

Compute L_k by pruning the itemset having minimum support count

K = K+1

Explanation

Consider the following database having the transactions <T1,T2,T3,T4,T5,T6,T7,T8> and assume that the minimum support is 50%. The database contains 4 items.

The minimum support is 50%. So the support count sup_count = 50/100*8 = 4

Table 2. Sample Database D

Transaction ID	Sequence of items
T1	3,2,4,3
T2	2,4,3
T3	4,2
T4	4,3
T5	1,2,1
T6	1,3,2
T7	2,4,3,1
T8	2,4,3

Step 1: Scan the database and identify the maximum length of the transaction δ.

Maximum length of the transaction δ = 4 (for Transactions T1 and T7)

Candidate itemset C_1={1,2,3,4}

Step 2: Frequent-1 itemset

Table 3. Frequent-1 Items

Item ID	Tid string	Count
2	T1, T2, T3, T5, T6, T7, T8	7
3	T1, T2, T4, T6, T7, T8	6
4	T1, T2, T3, T4, T7, T8	5

Frequent 1-itemset L_1={2,3,4}

Step 3: Construction of F-2 matrix by scanning the database.

Table 3. F-2 Matrix

	2	3	4
2	0	0	T1,T2,T7,T8
3	T1,T6	0	0
4	T3	T1,T2,T7,T8	0

Candidate itemset C_2 = { (2,4), (3,2), (4,2), (4,3) }

Step 4: Frequent Itemset for k=2.

Table 4. Frequent – 2 Itemsets

Item ID	Tid string	Count
2,4	T1,T2,T7,T8	4
4,3	T1,T2,T7,T8	4

Frequent 2-itemset L_2={ {2,4},{4,3} }

Step 5: Frequent Itemset for k=3

Iteration 1: Candidate itemset for k=3 from the matrix {common TID for (2,4) & (4,2) gives TID for (2,4,2)}.

Table 5. Candidate – 3 Itemsets

Item ID	Tid string	Count
2,4,2	Null	0
2,4,3	T1,T2,T7,T8	4
4,3,2	T1	1
4,3,4	Null	0

Candidate itemset for k=3 from the matrix are C_1= { {2,4,2},{2,4,3},{4,3,2},{4,3,4}}

Iteration 2:

Table 6. Frequent – 3 Itemsets

Item ID	Tid string	Count
2,4,3	T1,T2,T7,T8	4

Frequent 3-itemset L_3={ {2,4,3}}

Given all frequent itemsets, we can generate all frequent and confident association rules. For this first, all frequent itemsets are generated using Algorithm. The association rule can be mined from this frequent sequential pattern. Example, 50% of the users who visit the web page 2 and web page 3 also visit the web page 4. With this knowledge the web site can be restructured to better facilitate information retrieval.

V. EXPERIMENTAL RESULTS

In this section we performed a set of experiments to evaluate the effectiveness of the frequent sequential pattern mining using F-2 matrix method. The algorithm FSPM was executed on a Pentium 4 CPU, 2.4GHz, and 4 GB of RAM computer. The experiment database sources are msnbc dataset and CTI dataset, provided by the UCI data repository. The experiments have been conducted for the proposed algorithm (FSPM) and the results are compared with the algorithms such as Apriori, prefix span.

Figure 3 and Figure 4 represents the results of msnbc dataset. In Figure 3, the number of transactions is taken as 50,000 and the minimum support threshold is varied from 5% to 30% (x-axis). In Figure 4 the minimum support threshold is 25% and the size of the database (number of transactions) has been varied from 10,000 to 50,000(x-axis).

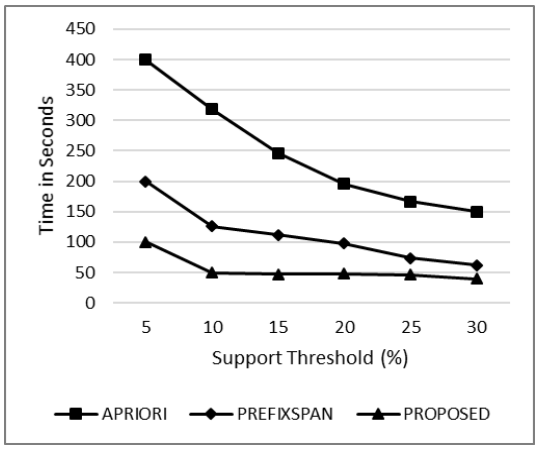

Figure 3. Results of msnbc dataset with varying min_sup

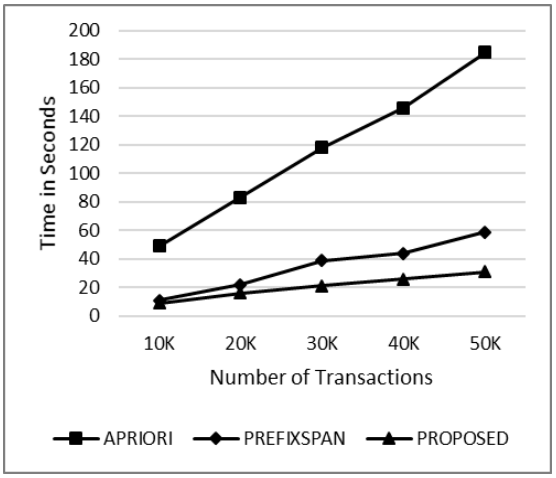

Figure 2. Results of msnbc dataset with varying database size

The execution time decreases when the support threshold increases and the execution time increases as the number of transaction increases. Thus from the analysis it is clear that the proposed method provides less execution time when compared with the other existing techniques for msnbc dataset.

Figure 5 and Figure 6 represents the results of CTI dataset. In Figure 5, the number of transactions is taken as 10,000 and the minimum support threshold is varied from 5% to 30% (x-axis). In Figure 6, the minimum support threshold is 25% and the size of the database (number of transactions) has been varied from 2,000 to 12,000 (x-axis).

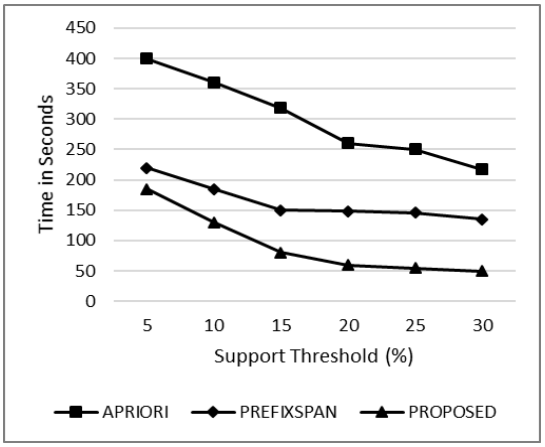

Figure 3. Results of CTI dataset with varying min_sup

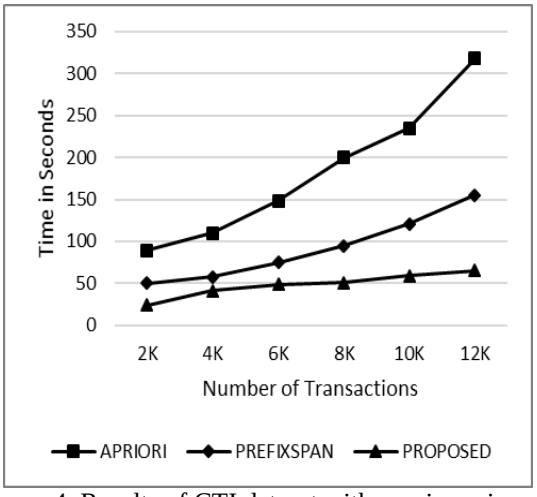

Figure 4. Results of CTI dataset with varying min_sup

Thus the result analysis in Figure 5 and Figure 6 shows that the proposed method provides less execution time when compared with the other existing techniques for CTI dataset.

VI. CONCLUSION AND FUTURE SCOPE

Determining frequent sequential Patterns is one of the most important fields of data mining. We presented a new, novel and efficient research approach on frequent sequential pattern mining using matrix representation. Its idea is to reduce the number of database scan, where the frequent items and the item id string is represented in matrix form. This approach can be used for the databases which the number of items is less and the number of transactions are high. The performance study shows that This approach mines the frequent sequential patterns efficiently and runs significantly faster than both Apriori and Prefixspan. This approach can be used to reduce the execution time and running out of large amount of memory and improving response time. As popularity of the web continues to increase, there is a growing need to develop tools and techniques that will help

to improve its overall efficiency. Also this work can be extended to mine the frequent sequential patterns with time constraints and other kinds of time-related knowledge.

REFERENCES

[1] Renáta Iváncsy, István Vajk, "Frequent Pattern Mining in Web Log Data", Acta Polytechnica Hungarica, Vol. 3, No. 1, 2006.

[2] S. K. Madria, S. S. Bhowmick, W. K. Ng, and E.-P. Lim, "Research issues in web data mining", Data Warehousing and Knowledge Discovery, pp. 303-312, 1999.

[3] J. Borges and M. Levene, "Data mining of user navigation patterns, WEBKDD, pp. 92-111, 1999.

[4] M. N. Garofalakis, R. Rastogi, S. Seshadri, and K. Shim, "Data mining and the web: Past, present and future", ACM CIKM'99 2nd Workshop on Web Information and Data Management (WIDM'99), Kansas City, Missouri, USA, C. Shahabi, Ed. ACM, pp. 43-47, 1999.

[5] S. Chakrabarti, "Data mining for hypertext: A tutorial survey", SIGKDD: SIGKDD Explorations: Newsletter of the Special Interest Group (SIG) on Knowledge Discovery and Data Mining, ACM, Vol. 1, No. 2, pp. 1-11, 2000.

[6] M. Eirinaki and M. Vazirgiannis, "Web mining for web personalization", ACM Trans. Inter. Tech., Vol. 3, No. 1, pp. 1-27, 2003.

[7] J. Pei, J. Han, B. Mortazavi-Asl, and H. Zhu, "Mining access patterns efficiently from web logs", PADKK '00: Proceedings of the 4th Pacific- Asia Conference on Knowledge Discovery and Data Mining, Current Issues and New Applications. London, UK: Springer-Verlag, pp. 396-407, 2000.

[8] Vinita Shrivastava, Neetesh Gupta, "Performance Improvement Of Web Usage Mining By Using Learning Based K-Mean Clustering", International Journal of Computer Science and its Applications.

[9] Craig Peter Oosthuizen, "Web Usage Mining of Organisational Web Sites", December 2005.

[10] B.Mortazavi-asl, "Discovering and Mining User Web-Page Traversal Patterns", Computer Science, Simon Fraser University, 1999.

[11] M. Géry and H.Haddad, "Evaluation of web usage mining approaches for user's next request prediction", Proc. 5th ACM International workshop on web information and data management, New Orleans, Louisiana, USA. pp 74-81. 7-8 November, 2003.

[12] R. Kosala, H. Blockeel, "Web Mining Research: A Survey", SIGKKD Explorations, vol. 2(1), July 2000.

[13] J. Srivastava, R. Cooley, M. Deshpande, P.-N. Tan, "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data", SIGKKD Explorations, vol.1, Jan 2000.

[14] J. Punin, M. Krishnamoorthy and M. Zaki, "Web usage mining: Languages and algorithms, Studies in Classification", Data Analysis, and Knowledge Organization. Springer-Verlag, 2001.

[15] M. S. Chen, J. S. Park, and P. S. Yu, "Data mining for path traversal patterns in a web environment", Sixteenth International Conference on Distributed Computing Systems, pp. 385-392, 1996.

[16] P. Batista, M. Ario, and J. Silva, "Mining web access logs of an on-line newspaper", 2002.

[17] O. R. Zaiane, M. Xin, and J. Han, "Discovering web access patterns and trends by applying olap and data mining technology on web logs", ADL '98: Proceedings of the Advances in Digital Libraries Conference. Washington, DC, USA: IEEE Computer Society, pp. 1-19, 1998.

[18] J. F. F. M. V. M. Li Shen, Ling Cheng and T. Steinberg, "Mining the most interesting web access associations", WebNet 2000-World Conference on the WWW and Internet, 2000, pp. 489-494.

[19] B. Jeudy and F. Rioult, "Database transposition for constrained closed pattern mining", Proceedings of Third International Workshop on Knowledge Discovery in Inductive Databases (KDID) co-located with ECML/PKDD, 2004.

[20] R. Agrawal, R. Srikant, "Fast algorithms for mining association rules", Proceedings of the 20th International Conference on Very Large Data Bases, 1994, pp. 487–499.

[21] J. Han, "Research challenges for data mining in science and engineering", NGDM 2007.

[22] R. Agrawal, R. Srikant, "Mining sequential patterns", Proceedings of the 11th International Conference on Data Engineering, pp. 3, 1995.

[23] Han J, Pei J, Yin Y, "Mining frequent patterns without candidate generation", Proceeding of the 2000 ACM-SIGMOD international conference on management of data (SIGMOD'00), Dallas, TX, pp 1–12, 2000.

[24] Jian Pei, Jiawei Han et al, "PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth", Proceeding ICDE '01 Proceedings of the 17th International Conference on Data Engineering, IEEE Computer Society Washington, DC, USA, 2001.

[25] Jiawei Han, Hong Cheng, Dong Xin, "Frequent pattern mining: current status and future directions", Data Mining and Knowledge Discovery, 15, 55–86, 2000.

[26] P.V. Nikam, D.S. Deshpande, "Different Approaches for Frequent Itemset Mining", International Journal of Scientific Research in Computer Science and Engineering, Vol.6, Issue.2, pp.10-14, 2018.

[27] Pradeep Chouksey, "*Mining Frequent model Using mass-produced Approach*", International Journal of Scientific Research in Computer Science and Engineering, Vol.5, Issue.4, pp.89-94, 2017.

[28] Sunil Joshi, "A Dynamic Approach for Frequent Pattern Mining Using Transposition of Database", The IEEE 2010 International Conference on Communication software and Netweorks (ICCSN 2010) from 26 - 28 February 2010.

[29] Sunil Joshi, Dr. R. S . Jadon, Dr. R. C. Jain, "An Implementation of Frequent Pattern Mining Algorithm using Dynamic Function", International Journal of Computer Applications (0975 – 8887), Volume 9– No.9, November 2010.

[30] Sonia Sharma, Munishwar Rai, "*Customer Behaviour Analysis using Web Usage Mining*", International Journal of Scientific Research in Computer Science and Engineering, Vol.5, Issue.6, pp.47-50, 2017.

