International Journal of Engineering Science Invention (IJESI)
ISSN (Online): 2319 – 6734, ISSN (Print): 2319 – 6726
www.ijesi.org ||Volume 7 Issue 1|| January 2018 || PP. 06-14

Efficient Approach for Finding Strong Patterns from Weblog using Web Usage Mining

[1]Ms.Karuna Nidhi Pandagre, [2]Dr S.Veenadhari
[1]Department of Computer Science
[2]Associate Professor, CSE Department Aisect University Bhopal, India
Research Scholar, Aisect University, Bhopal, India
Corresponding Author: [1]Ms.Karuna Nidhi Pandagre

Abstract: *Frequent pattern mining is one of the techniques of data mining. In this technique, generally set threshold value for finding strong patterns from the dataset or other types of repository. If the dataset is in the form of weblog then it is very challenging task to get strong patterns from these dataset because weblog contains semi-structured and unstructured information. Due to this, extracting strong patters or interesting patters from weblog is very challenging. Due to this problem web usage mining technique can be used. Through web usage mining techniques a visitor behavior can be predict and access intrinsic information from weblog. In this paper, we proposed an approach for extracting strong patterns from weblog. In this paper, we have proposed an algorithm named SPMW (Strong Patterns Mining from Weblog) algorithm and framework from mining strong patterns from weblog.*

Keywords: *Strong Patterns, Weblog, Web Usage Mining, Threshold Value*

Date of Submission: 15-11-2017 Date of acceptance: 23-12-2017

I. Introduction

The collection of minimum frequent itemset might be important. An example can be drowning in drug analysis, market basket analysis, business analysis, etc. Most of the criteria are based on support and confidence, here the support consists number of times pattern occur in the transaction databases , moreover it's a frequency of itemset in a transactional database and the confidence determines the proportion value that shows how frequently a part of the pattern (premise), occurs among all the records containing the whole transaction dataset. For example, if the pattern has to satisfy the minimum support then that pattern is considered as frequent pattern or frequent pattern on the contrary, these patterns have to satisfy maximum support then that pattern considered as in frequent pattern [1] [2]. Frequent patterns can be used in different domains such as biology, medicine and security [3][20], etc. For example, in a clinical database analysis one can discover frequent patterns that will help doctors to make decisions about the clinical care. As one can observe, each type of patterns expands the data seeking for specific types of knowledge. Moreover mining frequent itemsets are one of the most complicated and challenging tasks. The objective of Frequent Itemset Mining (FIM) is to discover important information as a pattern format from the huge amount of data [7][24]. Here the huge amount of data in form of weblog. There are following attributes of weblog are shown in figure 1.

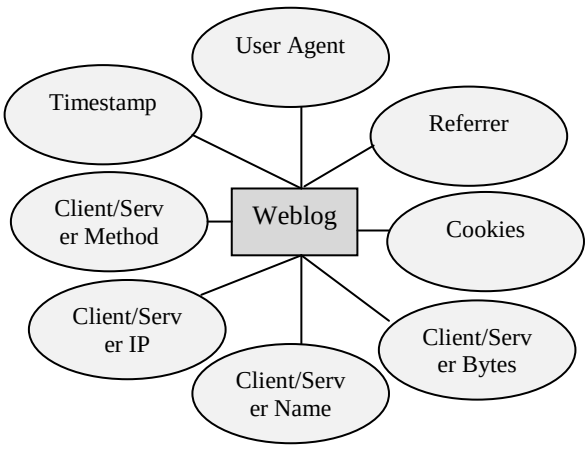

Figure1. Attributes of Weblog

Association Rule Mining (ARM) is one such technique, it aims to extract interesting correlations, frequent patterns, associations or casual structures among sets of items in the transaction databases or other data repositories . Apriori algorithm and Frequent Pattern Growth (FP-Growth) mining are the very famous algorithms of ARM. The performance point of view, Apriori algorithm is not very efficient as it demands database to be scanned again and again in turn increases the time complexity[22][23]. To overcome this problem, FP-Growth mining algorithm has been used.

In other types of patterns 'frequent' patterns that can be mined. Any item set is found interesting only when its frequency is less than the maximum threshold or more than the minimum threshold [5]. For searching 'frequent patterns' is an NP- Hard problem whose complexity is exponential [4]. This is complex from the computational point of view. A few algorithms have been developed which can search the frequent patterns in NP-Complete time or we can say it's solved such problems in polynomial time. Algorithms to search for both the patterns are rarely available apart from the 'Rarity' these are many such problems which exist in different data mining algorithms [6]. We have taken log data is collected which gets available at The Internet Traffic Archive sponsored by ACM SIGCOMM [9]. This log data later partitioned on the basis of its attributes and we have chosen two field timestamp and web page after applying preprocessing techniques [5] [4].

II. Web Mining

Web mining is the applications of data mining technique which is used to mine intrinsic information from weblog [9]. Whatever interesting data has to retrieve from Web, It is also possible through web mining [8]. Today huge amount of data is available on the web to extract data from such vast collection is a complex task. By applying some data mining method, we can find out useful pattern using web mining [8][19]. Web mining has been classified into three types: Web Structure Mining, Web Content Mining, Web Usage Mining.

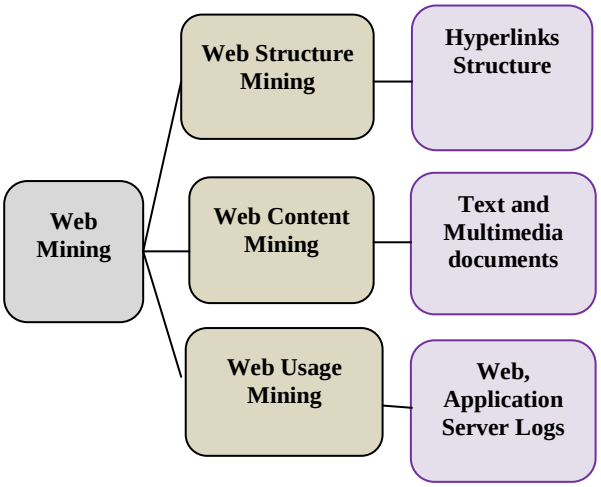

Figure2. Types of Web Mining

Web Structure Mining

Through Web Structure mining, useful information is extracted from hyperlink [9][18]. Which represent the structure of web, just like we can discover important page, we can analyze it. It is widely used in search engines. Through Web Structure mining we can discover community of such user who share common interest. In traditional Data mining technique such tasks are not performed because generally in relational table such link structure is not present.

Web Content Mining

Web Content Mining extracts useful information from web page contents. Through web content mining, we can automatically classify and cluster according to web page topics. This task is similar to data mining techniques. In web content mining we can deal with document in the form of text, audio, video, image and other contents.

Web Usage Mining

Web usage Mining is the application of Data Mining technique to discover user access pattern from web usage log [8][17]. It is used to provide better services of the website through which every click done by the user is measured. In figure 3, showing model of web usage mining

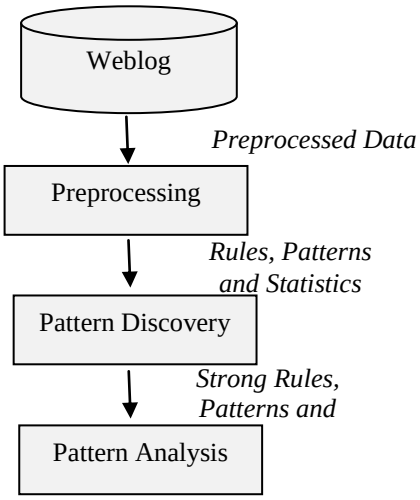

Figure3. Web Usage Mining Model

Log Preprocessing
It consists Data Cleansing, User Identification, Session Identification.
Pattern Discovery
It is a process to find out patterns in web logs but is frequently approved only on samples of data. Pattern Discovery Process are Statistical Analysis, Association Rules, Clustering, Classification, Sequential Patterns.
Statistical Analysis
By analyzing the log file, can be perform different kinds of statistical analysis like mean, median, mode, frequency etc.
Association Rules
It is refers to set of pages that are accessed together with a support value exceeding some specified threshold. Apriori and FP-Growth are association rule mining technique to find frequent itemsets.
Clustering
It is a technique to group together a set of items having similar characteristics. In Web Usage Mining two kinds of interesting cluster are discovered.
User Cluster
Establish groups of user having similar browsing pattern.
Page Cluster
Discover group of pages having related content.
Classification
It is a mapping to data item into one of several predefined classes.
Sequential Patterns
It is to find inter-session patterns such that presence of a set of items is followed by another item.
Web marketers can predict future visit patterns which will be helpful in placing advertisements aim at certain user groups.
Pattern Analysis
It is a method to filter out uninteresting rules or patterns from the set found in the pattern discovery phase.

III. Literature Review

Mining frequent pattern is one of the most research areas in various fields. Pyun G. et al. [11] propose a novel data structure named Linear Prefix (LP) tree. LP tree is composed of array forms and minimizes pointers between nodes and it is also uses minimum information required in the mining process and linearly accesses corresponding nodes. This data structure is also beneficial to the mining process. The algorithm is evaluated through various experiments, and the experimental results show that our approach outperforms previous algorithms in term of the runtime, memory, and scalability. Vijayalakshmi S. et al. [12] proposed a technique called AWAPT (Adaptive Web Access Pattern Tree). An AWAPT is a frequent sequence pattern technique for mining and combines Suffix tree and Prefix tree for efficient storage of all the sequences that contain a given item. They also proposed one of the data structure called WAP (Web Access Pattern) Tree, this data structure is for mining sequential pattern from the weblog. This WAP tree is much similar to Frequent Pattern (FP) Tree but

they reduced numerous reconstructions. The AWAPT algorithm is able to quickly determine the suffix of any frequent pattern prefix under consideration by comparing the assigned binary position codes of nodes of the tree. Sun L. et al. [13] proposed a new algorithm called CFPM (Combined Frequent Pattern Mining) for catering Web log data specifically. They used heuristics to prune the search space and reduce costs in mining so that better efficiency is achieved. They also studied the problem of mining Frequent Patterns (FP) from Web logs and done extensive experiments to examine the features of Web logs and the characteristics of FPs presented in such data. They also have conducted experiments to compare the performance of representative FP mining techniques on mining Web logs and have proposed improvements. Experiments show that the proposed algorithm CFPM significantly improves performance of the pattern growth approach.

Bing Liu et. al [14] proposed MSApriori algorithm to solve less frequent itemset problem because that a single minimum support is not sufficient for association rule mining since it cannot reflect the natures and frequency differences of the items in the database. MSApriori allows the user to specify multiple minimum supports to reflect the natures of the items and their varied frequencies in the database. Geng and Tian [15] present a new method to identify navigation related Web usability problems based on comparing actual and anticipated usage patterns. The actual usage patterns can be extracted from Web server logs. Through this they identified users, user sessions, and user task-oriented transactions and then they are applying a usage mining algorithm to discover patterns among actual usage paths. They also propose an approach to extract actual user behavior from Web server logs, capture anticipated user behavior with the help of cognitive user models and performs a comparison between the two [20][21].

IV. proposed approach

Most of mining attempt use to progress quality of data. To analyze user behavior, weblog data play a very important role for it. But weblog data contain some unnecessary information like image access, failed entries, server load etc [6][14]., which will affect the precision of pattern discovery and pattern analysis. Because of this data

Preprocessing is very important task in mining to find efficient patterns and getting efficient result.
Here, the log file is available at The Internet Traffic Archive sponsored by ACM SIGCOMM [25]. We use the part of the logs during the period of 1 July to 31 July1995. For session identification, set the maximum elapsed time to 30 min, which is used in many commercial applications [16].

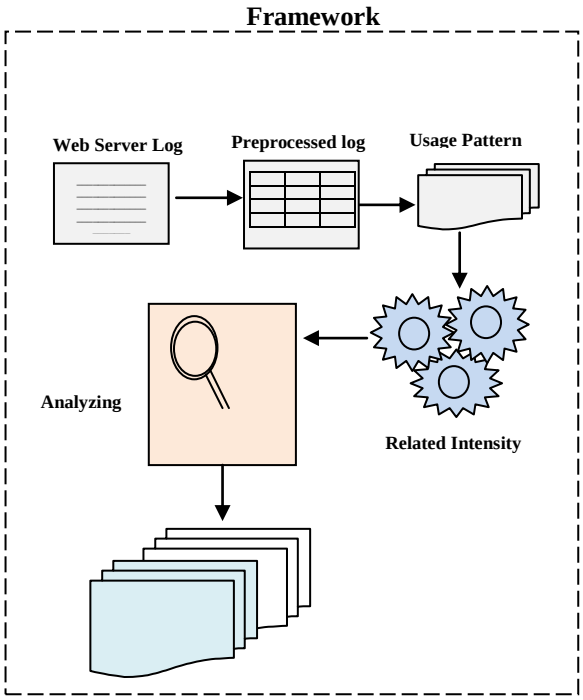

Figure4. Framework for Strong Patterns Mining

This is the first step to performed in the in the preprocessing of weblog data. It is the procedure to eliminate irrelevant information from weblog data. Lots of users can be accessed a website, but when HTTP status code failed with record it also maintain in weblog data [12][13]. Data cleaning is generally site specific procedure; lots of extra information may not be significant for analyzing the weblog data like style of file, sound

(audio, video) files [7]. Therefore some useless entries may occur and it can be remove in data cleaning process. Through data cleaning technique, inconsistency, irrelevancy, noise can be find out to improve the quality of data.

In this paper, Related Intensity (RI) value is calculated among visitor (V), web page (WP) and web browser (WB). There are various terms and definitions are used to calculate RI value.
Visitor (V) = {v_1, v_2, v_3.....v_n} is the set of visitor. Each web user can interact with the web server..

Web Page (WP) = {wp_1, wp_2, wp_3.....wp_n} is the set of web pages. The web pages contain information.

Web Browser (WB) = {wb_1, wb_2, wb_3..... wb_n} is the set of web browsers. A web browser is a software application for retrieving, presenting and traversing information resources on the WWW. There are various web browsers are available such as Internet Explorer, Opera, Mozilla Navigator Google Chrome etc.
Weight (W) = Weight has been used to find expected pattern. Through weight value can joined, visitor, web pages and web browsers. It is calculated by the following equations [16].

Table. I Relationships among Visitors, Web Pages and Web Browsers

V_{id}	WP_{id}	WP_{weight}	WB_{id}	WB_{weight}
1	1	5	1	3
1	2	4	2	5
2	3	3	4	7
2	1	2	1	5
2	2	7	3	18
3	2	6	3	7
3	4	5	5	18
4	1	6	2	2
4	2	8	1	6
5	3	3	5	3

According to table a graph can be created which contain three tuples (V ,WP, WB).In this every edge has a relation among different nodes and also having weight for every edge. Here Related Intensity Value (RIV) value can be calculated through (V, WP) (WP, WB). These relation are calculated using α_1 and α_1 Here the value α_1 and α_2 are predefined which must follow the relation $\alpha_1 + \alpha_2 = 1$. Here we have Taken $\alpha_1 = 0.6$ and $\alpha_2 = 0.4$.

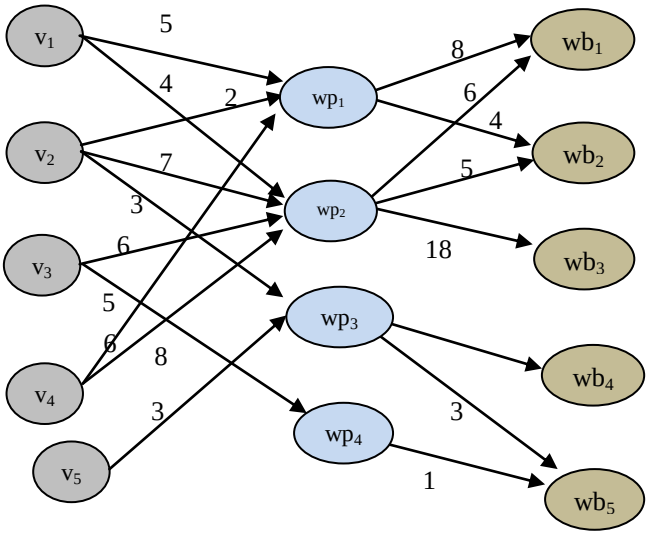

Figure5. Graph for V, WP and WB

By using this matrix, it can be created a above graph.

Relationships and RIV values using V1

As the above graph, we have Taken user V1, the calculation of relationships among user, time and image is depends on support of that relationships. Here we have assumed the minimum support is 0.3 .

$$\left[weight \text{ of } V \triangleright\triangleleft WP = \left[\frac{x1}{Max(x)}, \frac{x2}{Max(x)} ..... \frac{x3}{Max(x)} \right] \right]$$

(V1, wp1) = 5/5 = 1
(V1, wp2) = 4/5 = 0.8
(V1, wp3) = 1/5 = 0.2 (reject)

$$\left[weight \text{ of } WP \triangleright\triangleleft WB = \left[\frac{y1}{z1}, \frac{y2}{z2} ..... \frac{yn}{zn} \right] \right]$$

(wp1, wb1) =8/18 =0.44
(wp1, wb2) =4/18 =0.11 (reject)
(wp2, wb1) =6/18 =0.33
(wp2, wb2) =5/18 =0.27 (reject)
(wp2, wb3) =11/18 =0.61

In this above process if our value is lesser then the predefined support count then we can reject that pattern. Now RIV value can be calculated the following formula.(α_1 =0.6 and α_2 =0.4)

RIV= $[(E_1 \times (V, WP) \times \alpha_1) + (E_2 \times (WP, WB) \times \alpha_2)]$
V1 ⟶ wp1 ⟶ wb1
= $[1 \times 0.6 + 0.4 \times 0.4]$ = **0.76**
V1 ⟶ wp1 ⟶ wb2
= $[1 \times 0.6 + 0.4 \times 0.11]$ = **0.64**
V1 ⟶ wp1 ⟶ wb1 > V1 ⟶ wp1 ⟶ wb2

Relationships and RIV values using V2
(v2, wp1) = 2/7 = 0.28 (reject)
(v2, wp2) = 7/7 = 1
(v2, wp3) = 3/7 = 0.42
(wp2, wb2) =5/18 =0.27 (reject)
(wp2, wb3) =11/18 =0.61
(wp3, wb4) =7/18 =0.38
(wp3, wb5) =3/18 =0.16 (reject)

Now RIV Values of these Relations
RIV= $[(E_1 \times (V, WP) \times \alpha_1) + (E_2 \times (WP, WB) \times \alpha_2)]$
v2 ⟶ wp2 ⟶ wb3
= $[1 \times 0.6 + 0.61 \times 0.4]$ = **0.84**
v2 ⟶ wp3 ⟶ wb4
= $[0.4 \times 0.6 + 0.38 \times 0.4]$ = **0.39**
v2 ⟶ wp2 ⟶ wb3 > v2 ⟶ wp3 ⟶ wb4

Relationships and RIV values using V3
(v3, wp2) = 6/6 = 1
(v3, wp4) = 5/6 = 0.83
(wp2, wb1) =6/18 =0.33
(wp2, wb2) =5/18 =0.27 (reject)
(wp2, wb3) =11/18 =0.61
(wp4, wb5) =16/18 =0.88
v3 ⟶ wp2 ⟶ wb1

Now RIV Values of these Relations
RIV= $[(E_1 \times (V, WP) \times \alpha_1) + (E_2 \times (WP, WB) \times \alpha_2)]$
= $[1 \times 0.6 + 0.33 \times 0.4]$ = **0.73**
v3 ⟶ wp2 ⟶ wb3
= $[1 \times 0.6 + 0.61 \times 0.4]$ = **0.84**
v3 ⟶ wp4 ⟶ wb5
= $[0.83 \times 0.6 + 0.88 \times 0.4]$ = **0.85**
v3 ⟶ wp4 ⟶ wb5 > v3 ⟶ wp2 ⟶ wb3

> v3 ⟶ wp2 ⟶ wb1

Relationships and RIV values using V4
(v4, wp1) = 6/8 = 0.75
(v4, wp2) = 8/8 = 1
(wp1, wb1) =8/18 =0.44
(wp1, wb2) =4/18 =0.22 (reject)
(wp2, wb1) =6/18 =0.33
(wp2, wb2) =5/18 =0.27 (reject)
(wp2, wb3) =11/18 =0.61
Now RIV Values of these Relations
RIV= $[(E_1 \times (V, WP) \times \alpha_1) + (E_2 \times (WP, WB) \times \alpha_2)]$
v4 ⟶ wp1 ⟶ wb1
= $[0.75 \times 0.6 + 0.44 \times 0.4]$ = **0.62**
v4 ⟶ wp2 ⟶ wb1
= $[1 \times 0.6 + 0.33 \times 0.4]$ = **0.73**
v4 ⟶ wp2 ⟶ wb3
= $[1 \times 0.6 + 0.61 \times 0.4]$ = **0.84**
v4 ⟶ wp2 ⟶ wb3 > v4 ⟶ wp2 ⟶ wb1
> v4 ⟶ wp1 ⟶ wb1

Relationships and RIV values using V5
(v5, wp3) = 3/8 = 0.37
(wp3, wb4) =7/18 =0.38
(wp3, wb5) =3/18 =0.16 (reject)
Now RIV Values of these Relations
RIV= $[(E_1 \times (V, WP) \times \alpha_1) + (E_2 \times (WP, WB) \times \alpha_2)]$
v5 ⟶ wp3 ⟶ wb4
= $[0.37 \times 0.6 + 0.38 \times 0.4]$ = **0.37**

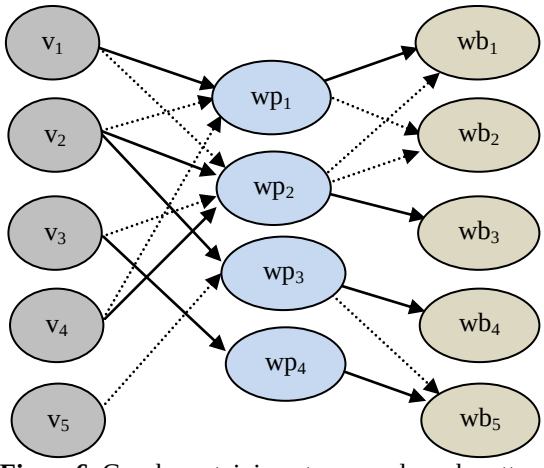

Figure6. Graph containing strong and weak patterns

Algorithm

SPMW (Strong Patterns Mining from Weblog) algorithm
General term using RIMWD: Weblog W, Visitor V, Web Page WP and Web Browser WB, Threshold Value TV
Input: Unstructured Dataset (UD)
Output: Strong Pattern Collection (SP)
1: Start
2: Take UD
3: Scan and UD
4: Read all the attributes of W and separate by space
5: If status code = Success
6: Then Take V, WP and WB
7: Calculate RIV from V, WP and WB see in Eq. 1, 2

8: Set TV
9: if RIV>=TV
10: then Accept those patters
11: Otherwise Reject those patterns
12: Assign all accepted patterns to SP List.
13: Stop

V. Experimental Analysis

In our experiment we have taken several web accessing patterns having different parameters V,WP and WB. We have provided different weight to each relation of these parameters. Then we find the strong patterns of the proposed approach and comparison is done with Apriori approach. In this comparison, we have analyzed that the proposed approach is better than the Apriori approach in terms of timing and large candidate datasets. Thus the proposed approach takes less time to generate strong frequent patterns.

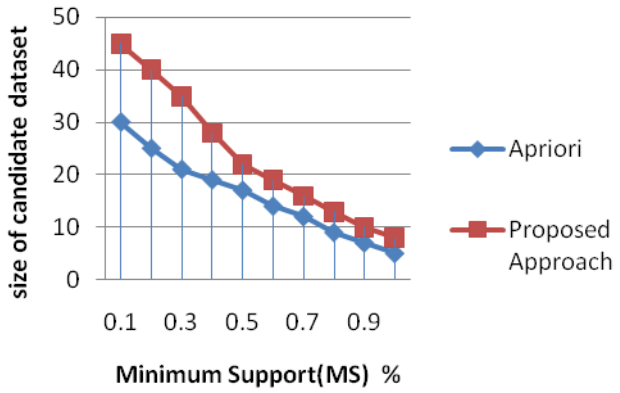

Figure7. Size of candidate dataset with different MS values

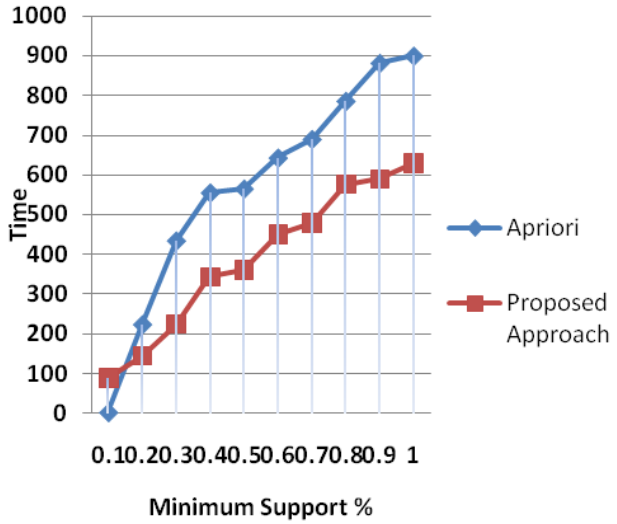

Figure8. Execution time with the different support values

VI. Conclusion

Now a days, large number of visitors are using internet services and internet service providers are available to provide services at free of cost such as social network services. business marketing services etc. This approach uses mapping visitor v, web pages wp and web browser wb. This approach generates strong patterns among these parameters, which is very useful to identify user in internet web service environment. The comparison of this approach is done with the standard Apriori algorithm, in which we find that the results of the pattern generation are excellent. The proposed technique including framework and algorithm are also useful in internet service providers. This is very useful approach for identifying a visitor behavior or activity from internet services. These approach are used in the different application of Web Usage Mining like E-Governance, E-Commerce and Education etc..

References

[1] Won Adda, W. Lei, W. Sharon and F. Yi,"Pattern Detection With Rare Item-Set Mining", International Journal on Soft Computing, Artificial Intelligence and Applications (IJSCAI), Vol.1, No.1, August 2012.

[2] E. Yafi, AS. Al-Hegami, MA. Alam, and R. Biswas, "YAMI: Incremental Mining of Interesting Association Patterns", International Arab Journal of Information Technology, pp. 504-510, 2012.

[3] R. Agrawal, T. Imielinski and A. Swami, "Mining association rules between sets of items in large databases", ACM SIGMOD International Conference on Management of Data, Vol 22 Issue 2,

[4] G. Yang, "The complexity of mining maximal frequent itemsets and maximal frequent patterns", In: KDD '04: Proceedings of 10th ACM SIGKDD International Conferences on Knowledge Discovery and Data Mining. New York, pp 344–353, 2004.

[5] B. Bakariya, K. Mohbey and G. Thakur, "An Inclusive Survey on Data Preprocessing Methods Used in Web Usage Mining", Proceedings of Seventh International Conference on Bio-Inspired Computing: Theories and Applications (BIC-TA 2012) Advances in Intelligent Systems and Computing, vol. 202, pp 407-416, 2013.

[6] T. Uno, T. Asai, Y. Uchida and H. Arimura, "LCM: An Efficient Algorithm for Enumerating Frequent Closed Itemsets". In: FIMI03: Proceedings of Workshop on Frequent Itemset Mining Implementations, vol 123, 2003.

[7] J. Han, J. Pei and Y. Yin, "Mining frequent patterns without candidate generation," in Proc. of ACM-SIGMOD 20th International Conference on Management of Data, 2000, pp. 1-12.

[8] R. Cooley, B. Mobasher and J. Srivastava, "Web Mining: Information and Pattern Discovery on the World Wide Web", Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence ICTAI, 1997.

[9] M. Gomes, Z. Gong, "Web Structure Mining: An Introduction", Proceedings of the 2005 IEEE International Conference on Information Acquisition, June 27 - July 3, 2005, Hong Kong and Macau, China.

[10] T. Aye: Web Log Cleaning for Mining of Web Usage Patterns, IEEE, 2011..

[11] G. Pyun, U. Yun and K. H. Ryu, "Efficient frequent pattern mining based on Linear Prefix tree," Knowledge-Based Systems, vol. 55, pp. 125-139, Jan. 2014.

[12] S.Vijayalakshmi, V.Mohan and S.Suresh Raja, "Mining of Users' Access Behaviour for Frequent Sequential Pattern from Web Logs," International Journal of Database Management Systems,vol.2, pp. 31-45, Aug. 2010.

[13] L. Sun and X. Zhang, "Efficient Frequent Pattern Mining on Web Logs," Advanced Web Technologies and Applications Lecture Notes in Computer Science, vol. 3007, pp 533-542, Apr., 2004.

[14] Liu B., Hsu W. and Ma Y., "Mining Association Rules with Multiple Minimum Supports", ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD-99), pp. 15-18, 1999.

[15] R. Geng and J. Tian, "Improving Web Navigation Usability by Comparing Actual and Anticipated Usage," IEEE transactions on human-machine systems, vol. 45, pp. 84-94, Feb. 2015.

[16] Chen T, Chou Y, Chen T (2012) Mining User Movement Behaviour Patterns in a Mobile Service Environment. IEEE Transactions on Systems, Man, and Cybernetics, 42:87-101. doi: 10.1109/TSMCA.2011.2159583.

[17] B. Bakariya, G.S. Thakur "Pattern Mining Approach for Social Network Service" National Academy Science Letters, Springer, DOI 10.1007/s40009-016-0529-y,Volume 40, Issue 3, June 2017. ISSN: 2250-1754 0.34

[18] B. Bakariya, G.S. Thakur, "Mining Rare Itemsets from Weblog Data" National Academy Science Letters, Springer DOI: 10.1007/s40009-016-0465-x, Volume-39, Issue-05, October- 2016.ISSN 0250-541X 0.34

[19] B. Bakariya, G.S. Thakur, "An Efficient Algorithm for Extracting High Utility Itemsets from Web Log Data" The Institution of Electronics and Telecommunication Engineers (IETE) Technical Review DOI: 10.1080/025646 02.2014.1000396, Volume-32, Issue-02, March- 2015. ISSN 0256-4602 1.304

[20] B. Bakariya, G.S. Thakur, "User Identification Framework in Social Network Services Environment", International Journal of Informatica Economică, DOI:10.12948/issn14531305/18.2.2014.02, Volume-18, Issue-02, Aug- 2014. ISSN 1453-1305.

[21] B. Bakariya, K. Chaturvedi, K. Singh, G. S. Thakur, "Efficient Approach for Mining Top-k Strong Patterns in Social Network Service", 5th International Conference on Eco-friendly Computing and Communication Systems (ICECCS-2016) Dec. 8-9, IEEE, MANIT Bhopal.

[22] K. Chaturvedi, R. Patel, DK Swami, " Fuzzy C-Means based Inference Mechanism for Association Rule Mining: A Clinical Data Mining Approach", Volume 6, Issue 6, Pages 103-110, 2015.

[23] K. Chaturvedi, R. Patel, DK Swami, "An Inference Mechanism Framework for Association Rule Mining", International Journal of Advanced Research in Artificial Intelligence (IJARAI), Volume 3, Issue 9, Pages 1-8, 2014.

[24] K. Chaturvedi, R. Patel, DK Swami, "An Efficient Binary to Decimal Conversion Approach for Discovering Frequent Patterns", International Journal of Computer Applications, Volume 75, Issue 12, 2010.

[25] http://ita.ee.lbl.gov/html/contrib/NASA-HTTP.html

International Journal of Engineering Science Invention (IJESI) is UGC approved Journal with Sl. No. 3822, Journal no. 43302.

Ms.Karuna Nidhi Pandagre. "Efficient Approach for Finding Strong Patterns from Weblog using Web Usage Mining." International Journal of Engineering Science Invention(IJESI), vol. 07, no. 01, 2018, pp. 06–14.