Discovery of Frequent Usage Pattern for Web Data to Optimized Web based Applications

Jaswinder Kaur
Research Scholar
Dept. of Computer Science & Applications,
Kurukshetra University, Kurukshetra, Haryana, India

Kanwal Garg, PhD
Assistant Professor
Dept. of Computer Science & Applications,
Kurukshetra University, Kurukshetra, Haryana, India

ABSTRACT
The traffic on WWW is increasing at a rapid rate due to users interaction with web sites, these activities contributes to enormous information which is maintained in Web log file. Web usage mining plays an important role in discovering frequent pattern from Web data, that helps to better serve the need of Web based applications. In present research work, researcher finds out different user and their session, which help in identifying unique user's navigational path from pre-processed Web log data. Further, researcher also proposed Modified Apriori algorithm which helps in extracting frequent usage pattern using average support.

Keywords
Frequent Pattern, Average Support, Web Usage Mining, Web data, Pre-processed

1. INTRODUCTION
Web log mining also known as Web Usage mining is application of mining techniques aimed to obtain interesting and frequent user access patterns from web log files, browser logs or proxy server logs. Web usage mining consist of three important phases which are preprocessing, pattern discovery and pattern analysis. In preprocessing, conditioning of data is carried out as real world Web data may contain anamolies which are to be removed to have data with useful information. During second phase, which is data mining, techniques are applied on pre-processed data to discover frequent usage patterns. Various techniqes used for pattern discovery are stastical analysis, clustering, classification, association rule and sequential pattern. In third phase,which is termed as pattern analysis, where obtained usage patterns are analyzed to filter irrelveant information and extract the valuable information [1].

In this paper, pre-processed web access log file of size 11,625 KB is used to discover frequent pattern and is oragnized as follows: section 2 presents related work which concerns in discovering frequent pattern from web data. section 3 describes the way that helps in finding user and session. Modified apriori algorithm is explained in section 4. Experiment results is discussed in section 5, Finally section 6 concludes the paper.

2. LITERATURE REVIEW
Pattern discovery is an essential phase in Web usage mining. At this stage, data mining techniques is applied on pre-processed Web log data in order to extract frequent user pattern. Veeramalai et.al. (2010)[2] proposed Enhanced Modified Apriori hash tree with fuzzy algorithm is used to overcome Crisp boundary problem. Kiruthika et.al. (2011)[3] applied Clustering approach on processed web log data and finds strong association rule. Rahul & Abha (2012)[4] presents FP-growth model which uses Fp-tree data structure to obtain frequent pattern from Web log data. Latheefa & Rohini (2013)[5] developed a Custom-Built Apriori tool to discover interesting frequent access pattern in Web usage data. Alagesh & Ramaraj (2013)[6] discussed the usage of frequent user access pattern and analyzed the gap between existing technology and requirement. Meera & Firoz (2014)[7] proposed hybird approach of FP-Growth algorithm and Decision Tree. FP-Growth algorithm is used to remove the unimportant information from the contents and Decision tree is used to fetch the contents from Web page. Aanum (2015)[8] analyzed two algorithms namely apriori and fp growth to determine association rules that occur in the Web log dataset. Aarti et.al. (2015)[9] Clustering is used to find common behavioral users then Apriori algorithm is applied on clustered data to find access pattern.

3. USER AND SESSION IDENTIFICATIOIN
User identification is the process of identifying each user accessing the WWW. Objective of identifying an user is to know the access characteristics so as personalised services are provided. Session is the collection of activities performed since an user have logged in till the instant the user is logged off. WUM techniques are employed to identify different user session from the web access log, as the amount of data to handle is huge with intricacy. This is a very complex process owing to the reason of presence of firewalls, cache and proxy server.

- There are incidents where different user have same IP address and same browser or same operating system, in such cases URI eneteries are taken into account for user identification.
- Time oriented approach, in which time between access request by an user is taken into consideration, if it surpass a limit of 30 miutes(default session timeout) or timestamp on two consecutive accesed pages of a user is more than 10 minutes, it implies a new user session have started.

4. MODIFIED APRIORI ALGORITHM
It searches for all frequent itemset using candidate generation from Web data.This algorithm follows a level-wise searching or breath first search using frequent Web pages. Those itemset above the average support are called frequent itemset.

Modified Apriori algorithm follows 2 phase:

- Generate Phase: In this phase candidate(k+1)-itemset is generated using k-itemset, this phase creates C_k candidate set.
- Prune Phase: In this phase candidate set is pruned to generate large frequent itemset using "average support" as the pruning parameter. This phase creates L_k large itemset [10].

5. EXPERIMENTAL RESULTS

On completion of data cleaning in SQL developer, processed Web data is as shown in Fig 1. Which is obtained by removing data containing various types of anomalies.

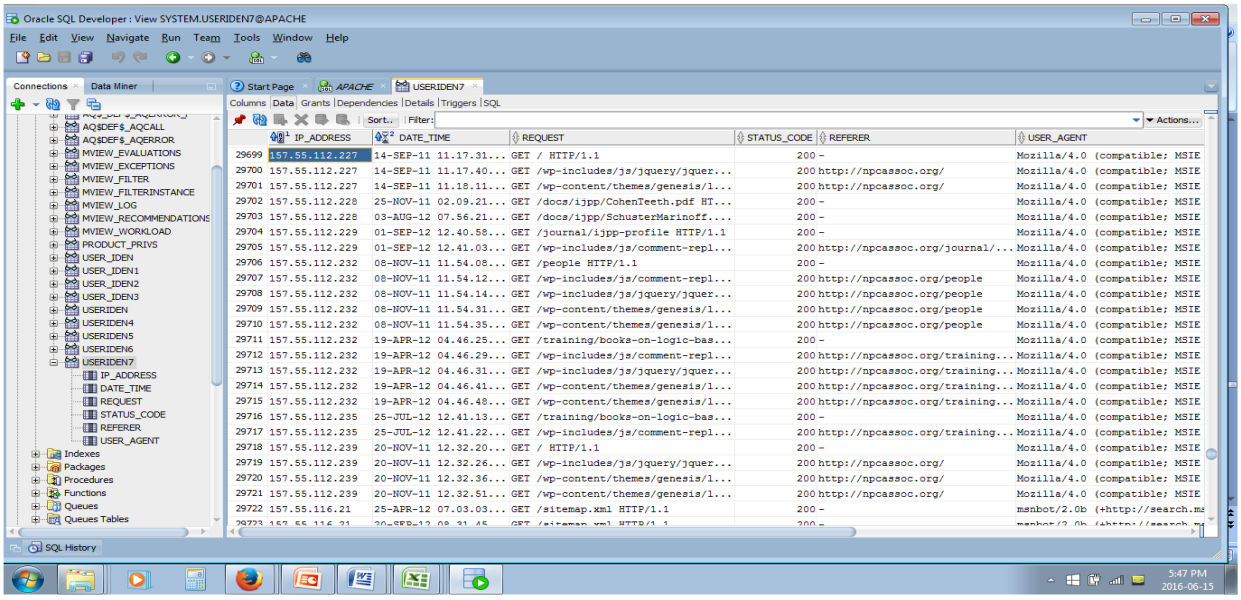
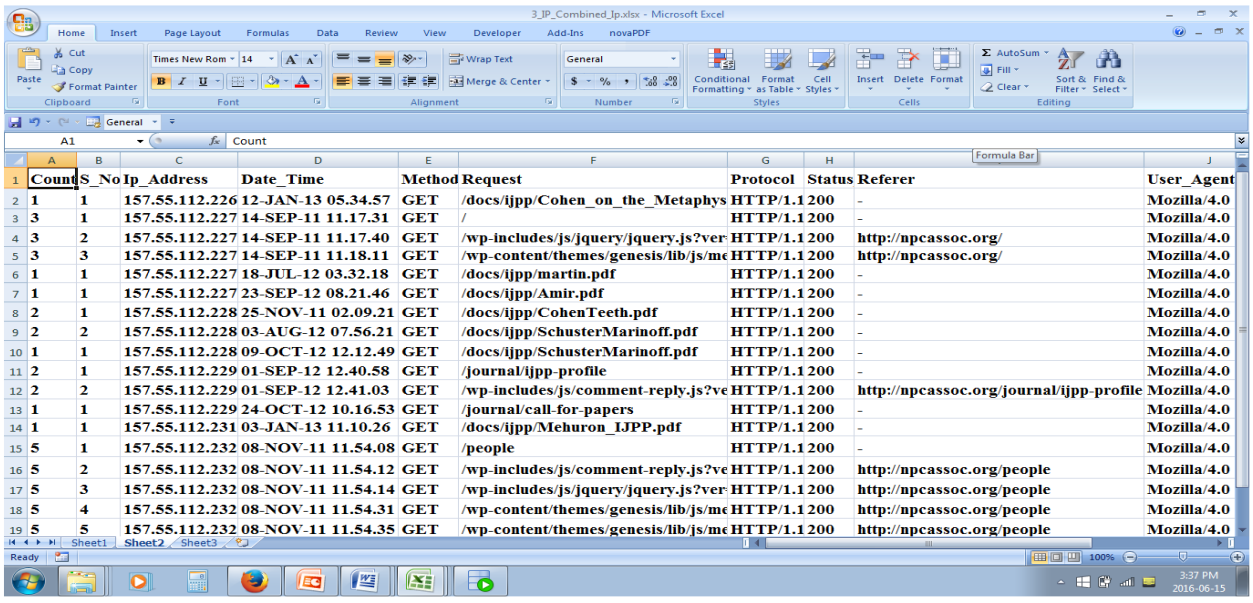


Fig 1: Cleaned Database Table

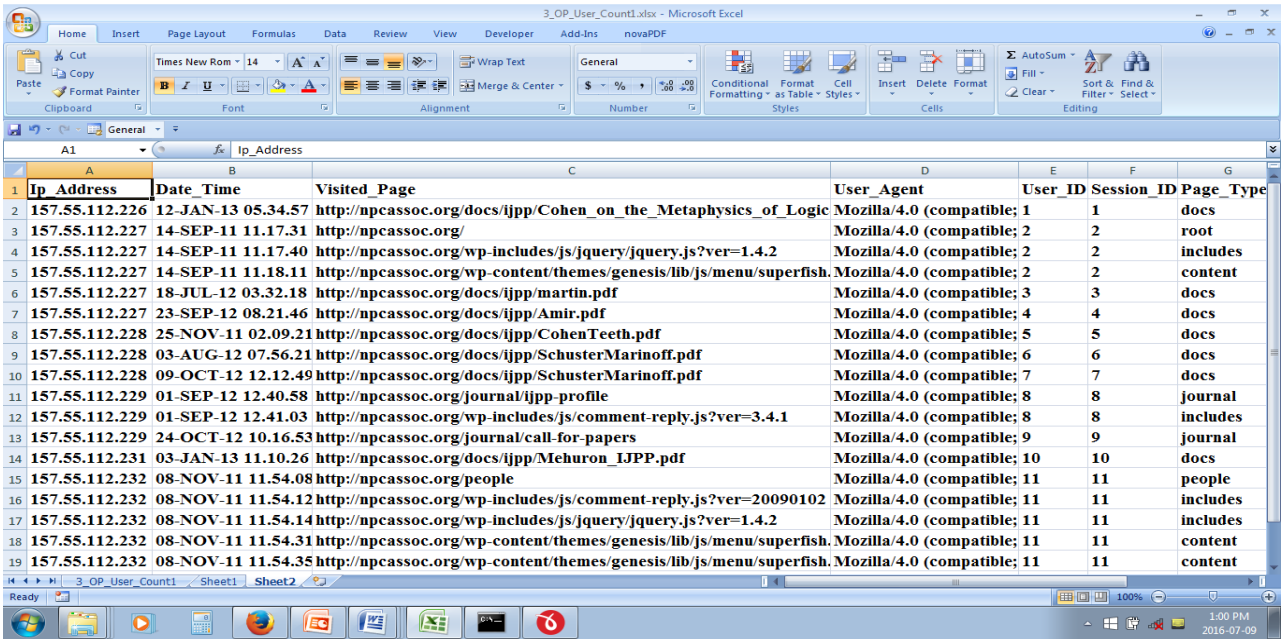
In Fig 2, user identification is done by using the navigational path for each user, different approach in user identification is done as proxy servers may assign same IP address to different users and these users may have similar user agent.



	A	B	C	D	E	F	G	H	I	J
1	Count	S_No	Ip_Address	Date_Time	Method	Request	Protocol	Status	Referer	User_Agent
2	1	1	157.55.112.226	12-JAN-13 05.34.57	GET	/docs/ijpp/Cohen_on_the_Metaphys	HTTP/1.1	200	-	Mozilla/4.0
3	3	1	157.55.112.227	14-SEP-11 11.17.31	GET	/	HTTP/1.1	200	-	Mozilla/4.0
4	3	2	157.55.112.227	14-SEP-11 11.17.40	GET	/wp-includes/js/jquery/jquery.js?ver	HTTP/1.1	200	http://npcassoc.org/	Mozilla/4.0
5	3	3	157.55.112.227	14-SEP-11 11.18.11	GET	/wp-content/themes/genesis/lib/js/me	HTTP/1.1	200	http://npcassoc.org/	Mozilla/4.0
6	1	1	157.55.112.227	18-JUL-12 03.32.18	GET	/docs/ijpp/martin.pdf	HTTP/1.1	200	-	Mozilla/4.0
7	1	1	157.55.112.227	23-SEP-12 08.21.46	GET	/docs/ijpp/Amir.pdf	HTTP/1.1	200	-	Mozilla/4.0
8	2	1	157.55.112.228	25-NOV-11 02.09.21	GET	/docs/ijpp/CohenTeeth.pdf	HTTP/1.1	200	-	Mozilla/4.0
9	2	2	157.55.112.228	03-AUG-12 07.56.21	GET	/docs/ijpp/SchusterMarinoff.pdf	HTTP/1.1	200	-	Mozilla/4.0
10	1	1	157.55.112.228	09-OCT-12 12.12.49	GET	/docs/ijpp/SchusterMarinoff.pdf	HTTP/1.1	200	-	Mozilla/4.0
11	2	1	157.55.112.229	01-SEP-12 12.40.58	GET	/journal/ijpp-profile	HTTP/1.1	200	-	Mozilla/4.0
12	2	2	157.55.112.229	01-SEP-12 12.41.03	GET	/wp-includes/js/comment-reply.js?ve	HTTP/1.1	200	http://npcassoc.org/journal/ijpp-profile	Mozilla/4.0
13	1	1	157.55.112.229	24-OCT-12 10.16.53	GET	/journal/call-for-papers	HTTP/1.1	200	-	Mozilla/4.0
14	1	1	157.55.112.231	03-JAN-13 11.10.26	GET	/docs/ijpp/Mehuron_IJPP.pdf	HTTP/1.1	200	-	Mozilla/4.0
15	5	1	157.55.112.232	08-NOV-11 11.54.08	GET	/people	HTTP/1.1	200	-	Mozilla/4.0
16	5	2	157.55.112.232	08-NOV-11 11.54.12	GET	/wp-includes/js/comment-reply.js?ve	HTTP/1.1	200	http://npcassoc.org/people	Mozilla/4.0
17	5	3	157.55.112.232	08-NOV-11 11.54.14	GET	/wp-includes/js/jquery/jquery.js?ver	HTTP/1.1	200	http://npcassoc.org/people	Mozilla/4.0
18	5	4	157.55.112.232	08-NOV-11 11.54.31	GET	/wp-content/themes/genesis/lib/js/me	HTTP/1.1	200	http://npcassoc.org/people	Mozilla/4.0
19	5	5	157.55.112.232	08-NOV-11 11.54.35	GET	/wp-content/themes/genesis/lib/js/me	HTTP/1.1	200	http://npcassoc.org/people	Mozilla/4.0

Fig 2: Sample Information From Request, Referer, User Agent logs

User identification approach result is as shown in Fig 3. Where the approach enables us to identify different user with similar IP address and same user agent, in the example below row (3-4-5) indicates a single user where the IP address and user agent were identical, row 8 and row 9 in spite of having similar IP address and User agents indicates two unique users. Furthermore in identifying different users, session of a user is also shown in Fig 3. Where the first user is having one session row (3-4-5), second user row 8 with one session and the third user row 9 also have one session.



Ip_Address	Date_Time	Visited_Page	User_Agent	User_ID	Session_ID	Page_Type
157.55.112.226	12-JAN-13 05.34.57	http://npcassoc.org/docs/ijpp/Cohen_on_the_Metaphysics_of_Logic	Mozilla/4.0 (compatible;	1	1	docs
157.55.112.227	14-SEP-11 11.17.31	http://npcassoc.org/	Mozilla/4.0 (compatible;	2	2	root
157.55.112.227	14-SEP-11 11.17.40	http://npcassoc.org/wp-includes/js/jquery/jquery.js?ver=1.4.2	Mozilla/4.0 (compatible;	2	2	includes
157.55.112.227	14-SEP-11 11.18.11	http://npcassoc.org/wp-content/themes/genesis/lib/js/menu/superfish.	Mozilla/4.0 (compatible;	2	2	content
157.55.112.227	18-JUL-12 03.32.18	http://npcassoc.org/docs/ijpp/martin.pdf	Mozilla/4.0 (compatible;	3	3	docs
157.55.112.227	23-SEP-12 08.21.46	http://npcassoc.org/docs/ijpp/Amir.pdf	Mozilla/4.0 (compatible;	4	4	docs
157.55.112.228	25-NOV-11 02.09.21	http://npcassoc.org/docs/ijpp/CohenTeeth.pdf	Mozilla/4.0 (compatible;	5	5	docs
157.55.112.228	03-AUG-12 07.56.21	http://npcassoc.org/docs/ijpp/SchusterMarinoff.pdf	Mozilla/4.0 (compatible;	6	6	docs
157.55.112.228	09-OCT-12 12.12.49	http://npcassoc.org/docs/ijpp/SchusterMarinoff.pdf	Mozilla/4.0 (compatible;	7	7	docs
157.55.112.229	01-SEP-12 12.40.58	http://npcassoc.org/journal/ijpp-profile	Mozilla/4.0 (compatible;	8	8	journal
157.55.112.229	01-SEP-12 12.41.03	http://npcassoc.org/wp-includes/js/comment-reply.js?ver=3.4.1	Mozilla/4.0 (compatible;	8	8	includes
157.55.112.229	24-OCT-12 10.16.53	http://npcassoc.org/journal/call-for-papers	Mozilla/4.0 (compatible;	9	9	journal
157.55.112.231	03-JAN-13 11.10.26	http://npcassoc.org/docs/ijpp/Mehuron_IJPP.pdf	Mozilla/4.0 (compatible;	10	10	docs
157.55.112.232	08-NOV-11 11.54.08	http://npcassoc.org/people	Mozilla/4.0 (compatible;	11	11	people
157.55.112.232	08-NOV-11 11.54.12	http://npcassoc.org/wp-includes/js/comment-reply.js?ver=20090102	Mozilla/4.0 (compatible;	11	11	includes
157.55.112.232	08-NOV-11 11.54.14	http://npcassoc.org/wp-includes/js/jquery/jquery.js?ver=1.4.2	Mozilla/4.0 (compatible;	11	11	includes
157.55.112.232	08-NOV-11 11.54.31	http://npcassoc.org/wp-content/themes/genesis/lib/js/menu/superfish.	Mozilla/4.0 (compatible;	11	11	content
157.55.112.232	08-NOV-11 11.54.35	http://npcassoc.org/wp-content/themes/genesis/lib/js/menu/superfish.	Mozilla/4.0 (compatible;	11	11	content

Fig 3: User And Session Identification

For easy understanding, only the top three frequent pattern in each path whose support value is greater than the average support is taken into consideration and is as shown in Table 1. Which clearly elaborates that when the navigational path is increased the corresponding average support decreases.

Table 1. Frequent Pattern Report

TOP THREE FREQUENT PATTERN IN EACH PATH	SUPPORT > AVG SUPPORT
FREQUENT MONO PATH	
/docs	29729
/normal	25537
/login	20029
FREQUENT DI PATH	
/normal -> /docs	12971
/normal -> /root	7644
/docs -> /docs	6207
FREQUENT TRIA PATH	
/normal -> /docs -> /docs	3169
/normal -> /root -> / root	2530
/journal -> /includes -> /content	2385
FREQUENT TETTARA PATH	
/normal -> /journal -> /includes -> /content	1450
/journal -> /includes ->/includes -> /content	1190
/journal, -> /includes ->/content ->/content	1182

Comparison chart between average support approach and minimum support approach, helps in understanding that in average support approach when the pattern length increases the number of pairs keeps on increasing. Whereas in minimum support approach when the pattern length is increased the number of pairs decreases.

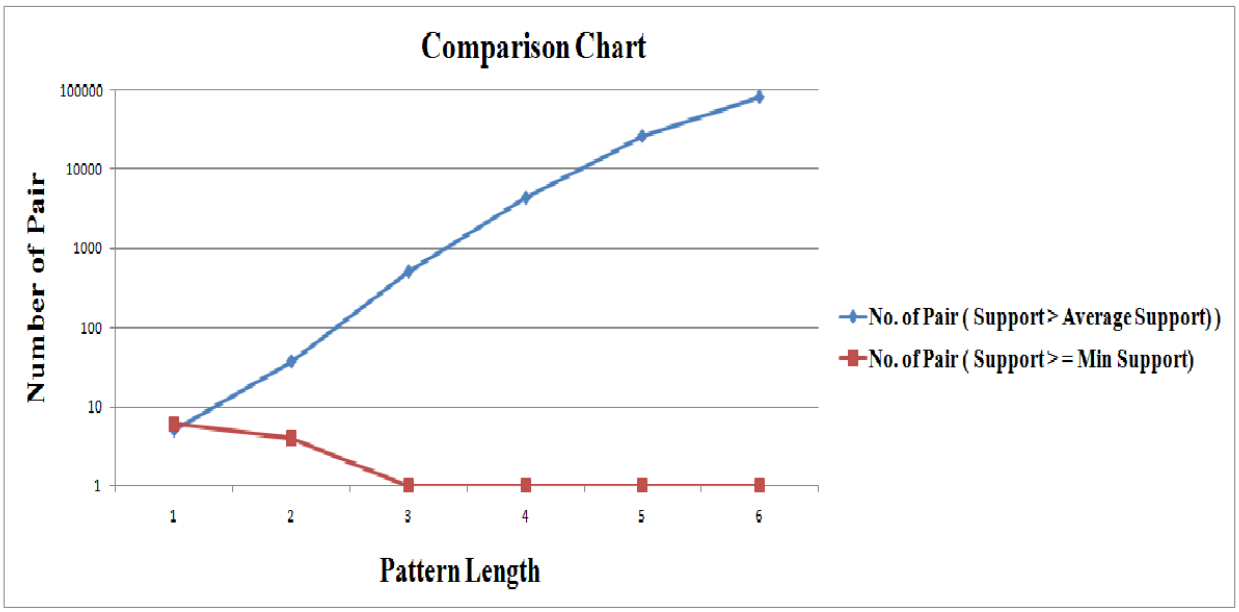


Fig 4: Comparison of Minimum Support And Average Support

6. CONCLUSION

The concept of web usage mining is easy to identify user's behaviour and interest, which helps web designer to optimized accessibility and usability of their websites. To need this idea researcher proposed Modified Apriori algorithim for discovering frequent pattern from Web log data by introducing average support value approach. This approach proves to be more efficient than already existing Apriori algorithm, as proposed algorithm generate frequent pattern with large pattern length, especially if large number of users exist in web log file. Proposed work can be implemented with more advance techniques/tools of web usage mining which will further assist in research field.

7. REFERENCES

[1] M. Valera, K. Rathod, "A novel approach of mining frequent sequential pattern from customized web log preprocessing", International Journal of Engineering Research and Applications, ISSN:2248-9622, Vol 3, Issue 1, pp.269-380, February 2013.

[2] S. Veeramalai, N. Jaisankar, A. kannan, "Efficient web log mining using enhanced apriori algorithm with hash tree and fuzzy", International Journal of Computer Science & Information Technology, Vol. 2, No. 4, pp. 60-74, August 2010.

[3] Ms K. M, Mr R. Jadhav, Ms R. J, Ms D. Dixit, Ms A. Nehete, Ms T. Khodkar, "Pattern discovery using Association rules", International Journal of Advanced Computer Science and Applications, Vol 2, No. 12, pp. 69-74, 2011.

[4] Mr. R. Mishra, Ms. A. Choubey, "Discovery of frequent patterns from web log data by using Fp-Growth algorithm for web usage mining", International Journal of Advanced Research in Computer Science and Software Engineering, ISSN: 2277 128X, Vol 2, Issue 9, pp.311-318, September 2012.

[5] L. V , R. V , "Web mining patterns discovery and analysis using custom-Built apriori algorithm", International Journal of Engineering Inventions, e-ISSN: 2278-7461, p-ISSN: 2319-6491, Vol 2, Issue 5, pp. 16-21, March 2013.

[6] P. A. Kannan, Dr. E. Ramaraj, " Usage and research challenges in the area of frequent pattern in data mining", ISRO Journal of Computer Engineering, e-ISSN: 2278-0661, p-ISSN: 2278-8727, Vol 13, Issue 2, pp. 08- 13, August 2013.

[7] M. J. Tank, F. A. Sherashiya, "Improved technique to discover frequent pattern using Fp-Growth and Decision Tree", International Journal of Engineering Development and Research, ISSN: 2321-9939, Vol 2, Issue 4, PP. 3551-3554, 2014.

[8] A. Shaikh, "Web usage mining using Apriori and Fp Growth algorithm", International Journal of Computer Science and Information Technologies, ISSN: 0975-9646, Vol 6, Issue 1, pp.354-357, 2015.

[9] A. M. Parekh, A. S. Patel, S. J. Parmar, Prof. V. R. Patel "Web usage mining: frequent pattern generation using association rule mining and clustering", International Journal of Engineering Research & Technology, ISSN: 2278-0181, Vol 4, Issue 04, pp. 1243-1246, April 2015.

[10] A. Saxena, S. Gadhiya, "A survey on frequent pattern mining methods Apriori, Eclat, Fp Growth", International Journal of Engineering Development and Research, ISSN: 2321-9939, Vol 2, Issue 1, pp. 92-96, 2014.