Sādhanā Vol. 40, Part 8, December 2015, pp. 2353–2365. © Indian Academy of Sciences

Prediction of users webpage access behaviour using association rule mining

R GEETHARAMANI[1], P REVATHY[2] and SHOMONA G JACOB[3,*]

[1]CEG Campus, Anna University, Chennai 600025, India
[2]Department of Computer Science & Engineering, Rajalakshmi Engineering College, Chennai 602105, India
[3]Department of Computer Science & Engineering, SSN College of Engineering, Chennai 603110, India
e-mail: rgeetha@auist.net; revathy.p@rajalakshmi.edu.in; shomonagj@ssn.edu.in

MS received 15 July 2014; revised 6 February 2015; accepted 15 June 2015

Abstract. Web Usage mining is a technique used to identify the user needs from the web log. Discovering hidden patterns from the logs is an upcoming research area. Association rules play an important role in many web mining applications to detect interesting patterns. However, it generates enormous rules that cause researchers to spend ample time and expertise to discover the really interesting ones. This paper works on the server logs from the MSNBC dataset for the month of September 1999. This research aims at predicting the probable subsequent page in the usage of web pages listed in this data based on their navigating behaviour by using Apriori prefix tree (PT) algorithm. The generated rules were ranked based on the support, confidence and lift evaluation measures. The final predictions revealed that the interestingness of pages mainly depended on the support and lift measure whereas confidence assumed a uniform value among all the pages. It proved that the system guaranteed 100% confidence with the support of 1.3E−05. It revealed that the pages such as Front page, On-air, News, Sports and BBS attracted more interested subsequent users compared to Travel, MSN-News and MSN-Sports which were of less interest.

Keywords. Apriori algorithm; association rules; data mining; MSNBC; web usage mining.

1. Introduction

Interest in the analysis of user behaviour on the Web has been consistently increasing (Chakrabarti 2002). This increase stems from the realization that added value for Website visitors are not gained merely through larger quantities of data on a site, but through easier access to the required information at the right time and in the most suitable form. Estimates of Web usage

*For correspondence

expect the number of users to be on the rise as their site activities show great improvement (Hung *et al* 2013; Phoa & Sanchez 2013; Anitha & Krishnan 2011). The latest statistics reveal the toll of Internet users to be nearly 1,386,188,112 in Asia and 3,035,749,340 around the world.

Majority of these users are non-expert and find it difficult to keep up with the rapid development of computer technologies, whereas at the same time they recognize that the Web is an invaluable source of information for their everyday life. The increasing usage of the Web also accelerates the pace at which information becomes available online. In various surveys of the Web (Eirinaki *et al* 2005; Jaideep *et al* 2000; Chifu & Salomie 2009; Debahuti 2010), e.g., it is estimated that roughly one million new pages are added every day and over 600 GB of pages change per month. A new Web server providing Web pages is emerging every two hours. Nowadays, more than three billion Web pages are available online, almost one page for every two people on the earth. Hence utilization of computational techniques to analyze and predict the interestingness of pages that could yield good prediction of possible website access was the motivation for this research (Chandra & Basker 2000; Sanchez & Liu 2011; Liu & Peng 2013).

Web Usage Mining is the application of data mining techniques to discover interesting usage patterns from Web data, in order to understand and better serve the needs of Web-based applications (Robert *et al* 1999; Suraya *et al* 2011; Jacob *et al* 2013). Data mining refers to the computational process of discovering patterns in large datasets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems (Kotsiantis *et al* 2007; Han & Kamber 2011; Ramani & Jacob 2013a, b; Jacob & Ramani 2013). Web Usage data captures the identity or origin of Web users along with their browsing behaviour at a Web site (Chun-sheng and Li 2014). Web usage mining itself can be classified further depending on the kind of usage data considered: Web Server Log: The user logs are collected by Web server. Commercial application servers such as Web logic Story Server have significant features to enable E- commerce applications to be built on top of them with little effort (Madhuri 2002; Yang *et al* 2009; Wen-Hai 2010; Veeramalai *et al* 2010). Web usage mining effectively has many advantages which makes this technology attractive to corporations including the government agencies.

In data mining, association rule mining (ARM) is a popular and well researched method for discovering interesting relations between variables in large databases (Agrawal & Srikant 1994; Kotsiantis & Kanellopoulos 2001; Kumar & Chezian 2012). It is intended to identify strong rules discovered in databases using different measures of interestingness. Apriori algorithm is the best-known association rule mining algorithm to mine association rules. It uses a breadth-first search strategy to count the support of itemsets and uses a candidate generation function which exploits the downward closure property of support (Wang *et al* 2000; Renáta & Vajk 2006). This research investigates the role of Apriori algorithm in mining association rules that unearth the interestingness of web pages from the MSNBC dataset.

2. State-of-the-art

A review of the recent and related work on Web usage mining is concisely presented in the following section.

Zhang & Chen (2014) presented a forensics method of web browsing behaviour based on association rule mining. The method aimed at providing the necessary data support to build the behaviour pattern library for investigation. The obtained original data was pre-treated to transactional data. Frequent browsing time and frequent web browsing sequences were obtained from the transactional data by Apriori algorithm. The mining results proved helpful to identify and recognize anonymous or suspicious web browsing behaviour patterns.

Tassa (2014) proposed a protocol for secure mining of association rules in horizontally distributed databases. The main ingredients in their protocol were two novel secure multi-party algorithms – one that computed the union of private subsets that each of the interacting players hold, and another that tested the inclusion of an element held by one player in a subset held by another. The proposed protocol offered enhanced privacy with respect to the protocol in (Zhang & Chen 2014). In addition, it was simpler and significantly more efficient in terms of communication rounds, communication cost and computational cost.

Mary & Malarvizhi (2012) proposed a new web page recommendation system based on weighted Apriori with dynamic programming. The dataset used in the study was MSNBC dataset. The authors extended association rule mining by assigning significant weights to the pages based on visiting order of each page. The system performance was evaluated based on time and space. The experimental result proved that the method guaranteed 35% support with 64% confidence, which was better than the conventional association rule models.

Sanchez & Liu (2011) used a Bayesian hierarchical model of the page counts per user to obtain posterior distributions of page access frequency that allow us to cluster user sessions in a relatively small number of groups. The proposed approaches were then generalized to different types of Web sites, different levels of aggregation of pages and different clustering schemes. Suraya *et al* (2011) suggested a personalized minimum support (P_minsup) threshold with user specified minimum items or (min_i). The authors used MSNBC dataset consisting of 989,818 users and 17 URL page categories. SPADE algorithm using vector array as an extension from the previous method of using relational database and pre-defined threshold was applied to this dataset. The result revealed that the performance of P_minsup (0.09%) with the complement of min_i value approach was effective based on the execution time in discovering frequent pattern in each k-sequence.

Suresh *et al* (2011) suggested that if data is too large with ample noise then the clustering method becomes very sensitive to find the initial center values. Improved fuzzy c-Means (FCM) clustering was used in finding the user access patterns from web access log. They used MSNBC web navigation dataset for forming the web data Clusters. Their results revealed that each of the clusters contained observations with specific common characteristics and improved the algorithm efficiency by identifying the initial cluster centers.

Santhisree & Damodaram (2010) presented an algorithm named OPTICS ("Ordering Points To Identify the Clustering Structure") to find density based clusters on the web usage data from the MSNBC.COM website. The average of inter cluster and intra cluster distance was calculated and the results were then compared with different similarity measures like Euclidean, Jaccard, projected Euclidean, cosine and fuzzy similarity to find the similarity between clusters and the results were visualized graphically to predict the user behaviour. Renáta & Vajk (2006) discovered hidden information from large amount of web log data collected by web servers. Their research introduced the process of web log mining, and showed how frequent pattern discovery tasks can be applied on the web log data in order to obtain useful information about the user's navigation behaviour.

The following section elaborates on the association rule mining framework proposed in this paper to explore the web site page's interestingness.

3. Association rule mining

Association rules are employed today in many application areas, an area of specific interest being Web usage mining (Agrawal & Srikant 1994; Kriegel 2007). Association rules are usually

required to satisfy a user-specified minimum support and a user-specified minimum confidence simultaneously (Kum *et al* 2005; Hacibeyoglu *et al* 2013; Ganapathy *et al* 2014; Zhou & Huang 2014). Association rule generation is usually split up into two separate steps: First, the minimum support being applied to find all frequent itemsets in a database. Second, the frequent itemsets and the minimum confidence constraint are used to form rules. Finding all frequent itemsets in a database is difficult since it involves searching all possible itemsets (item combinations). Efficient search is possible using the downward-closure property of support, which guarantees that for a frequent itemset, all its subsets are also frequent and thus for an infrequent itemset, all its supersets must also be infrequent. Exploiting this property, efficient algorithms (Apriori, Eclat) have been utilized in the past to discover frequent itemsets. The proposed framework is depicted below (figure 1).

3.1 *Data description*

The data is taken from Internet Information Server (IIS) logs for msnbc.com and news-related portions of msn.com for the entire day of September 28, 1999 (Gao *et al* 2009). Each sequence in the dataset corresponds to page views of a user and each event in the sequence corresponds to a user's request for a page. Requests are recorded at the level of page category (as determined by a site administrator). The categories are "frontpage", "news", "tech", "local", "opinion", "on-air", "misc", "weather", "health", "living", "business", "sports", "summary", "bbs" (bulletin board service), "travel", "msn-news", and "msn-sports". The data comprises 989,818 records with the 17 web pages being considered as the attributes.

3.2 *Data pre-processing*

The training data is downloaded from the UCI Machine Learning Repository and is stored as an Excel spreadsheet. The users are numbered from 1 through 989,818 forming the rows of the excel file whereas the pages visited by them form the columns numbered 1 till 17. The cells pertaining to each row and column provide information on the number of times (hits) each user visits each page. Once the data is saved in MS-Excel format, it is then loaded onto TANAGRA,

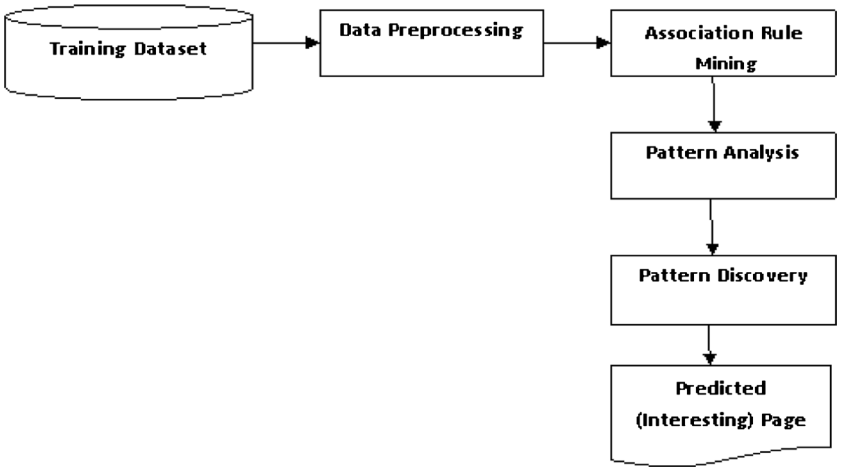

Figure 1. Framework of web access prediction.

open-source data mining suite. The data is visualized to guarantee proper loading and precise values of data storage following which association rule mining is applied.

3.3 *Apriori algorithm*

The Apriori algorithm is an influential algorithm for mining frequent itemsets for Boolean association rules. Detecting frequent itemsets involves the sets of items that have minimum support. The basic principle behind the Apriori algorithm states that 'A subset of a frequent itemset must also be a frequent itemset'. Hence if {AB} is a frequent itemset then both {A} and {B} should also be frequent itemsets (Jacob & Ramani 2012; Ramani *et al* 2012). In this research Apriori-prefix tree (PT) was utilized to generate the association rules with minimum support and confidence. Apriori-PT is a very powerful association rule generator, which can handle huge datasets very fast. At the time of execution, a temporary file is created and transmitted to the Apriori program from which the rules are automatically downloaded and displayed. Two major operations take place in Apriori execution. The Join Operation finds L_k, a set of candidate k-itemsets by joining L_{k-1} with itself. The prune step identifies (k−1)-itemset that is not frequent and prunes it.

The pseudocode of the Apriori algorithm is given below.

```
Ck: Candidate itemset of size k
Lk: frequent itemset of size k
L1= {frequent items};


For (k= 1; Lk!=∅; k++) do begin


Ck+1= candidates generated from Lk;
for each transaction tin database do increment the count of
all candidates in Ck+1that are contained in t
Lk+1= candidates in Ck+1 with min_support
end


return ∪kLk;
```

Execution of the Apriori algorithm on the MSNBC dataset generated all possible rules that satisfied the minimum support and confidence of 0.6% and 100% respectively. The maximum cardinal number of itemsets is 17. These parameters enabled to restrict the number of rules generated. These rules were then analyzed as discussed in the following section.

3.4 *Pattern analysis*

The patterns of rules generated were categorized based on the sequence of pages visited and the number of items (pages) in the sequence. For example P1∧P2∧P3 indicate a three-sequence itemset whereas p2∧p4∧p5∧p8 indicate a four-sequence itemset. The 3-sequence itemsets were considered as the threshold for the sequence number since two-sequence page visits carry the possibility of uninterested users. The number of rules generated for each sequence itemset is tabulated in table 1.

Table 1. Rules and sequence itemsets based on rule antecedents.

S. no	Sequence itemset	Number of rules	Min_Supp
1	3	38	0.0000152
2	4	922	0.0000495
3	5	7403	0.0000192
4	6	28281	0.0000283
5	7	64326	0.0001051
6	8	99266	0.0000626
7	9	103134	0.0000192
8	10	83186	0.0000778
9	11	49799	0.0000505
10	12	22093	0.0000495
11	13	7120	0.0000232
12	14	1584	0.0000141
13	15	218	0.0000131
14	16	5	0.0000131
15	17	9	0.0000131

Since the numbers of rules generated were increasing with the number of sequence item sets, a threshold support was fixed for each sequence as indicated in table 2. The above-mentioned approach was adopted to retrieve the rules considering the antecedent sequence. From the perspective of the consequents, each website page was analyzed individually to detect the frequency of visits for that particular page. The manner in which the interesting patterns were discovered is discussed in the following section.

4. Experimental results

The results are discussed in three sections. The first section briefs about the interestingness measures for an association rule. The second reports the discovered patterns from the rule antecedents while the final section reports on the patterns detected from the rule consequents.

Table 2. Rules and sequence itemsets based on rule consequents.

S. no	Pages	Number of rules	Lift
1	P1	4608	0.316
2	P6	4608	0.219
3	P2	4608	0.177
4	P3	3072	0.123
5	P4	4608	0.122
6	P14	4608	0.12
7	P12	4608	0.113
8	P8	4608	0.096
9	P9	3072	0.911
10	P7	3072	0.081
11	P13	3072	0.077
12	P11	4608	0.058
13	P10	4608	0.051
14	P5	4608	0.025

Table 3. Pattern discovery from rule antecedents in 3-sequence itemsets.

Rule no.	Rule	Highest lift	Predicted subsequent page
R1	P16∧P17∧P13	0.316402	P1
R2	P16∧P17∧P15	0.219335	P6
R3	P16∧P17∧P8	0.316402	P1
R4	P16∧P17∧P9	0.316402	P1

4.1 *Association rule measures*

The measures of interestingness that rank the association rules generated by any association rule mining algorithm include Support, Confidence and Lift. They are defined as follows.

(a) Support

The support of an itemset (X) is defined as the proportion of transactions in the dataset which contain the itemset.

(b) Confidence

The confidence of a rule is defined as

$$\text{conf}(X \Rightarrow Y) = \text{supp}(X \cup Y)/\text{supp}(X) \tag{1}$$

(c) Lift

The lift of a rule (x->y) is defined as the ratio of the observed support to that expected if X and Y were independent

$$\text{lift}(X \Rightarrow Y) = \frac{\text{supp}(X \cup Y)}{\text{supp}(X) \times \text{supp}(Y)} \tag{2}$$

The above measures were utilized to filter the irrelevant rules following which the relevant rule antecedents and rule consequents were analyzed.

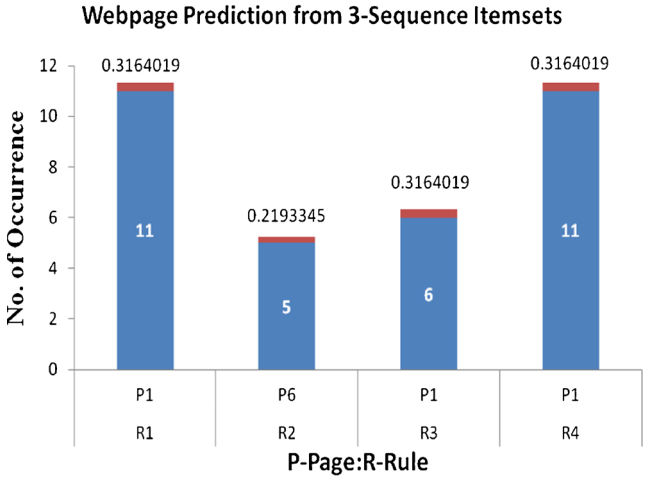

Figure 2. Webpage prediction from 3-sequence itemsets.

Table 4. Pattern discovery from rule antecedents in 4-sequence itemsets.

Rule no.	Rule	Highest lift	Predicted subsequent page
R1	P16∧P17∧P15∧P10	0.316	P1
R2	P16∧P17∧P15∧P8	0.316	P1
R3	P16∧P17∧P15∧P14	0.316	P1
R4	P16∧P17∧P15∧P3	0.219	P6
R5	P16∧P17∧P15∧P1	0.219	P6
R6	P16∧P17∧P5∧P7	0.316	P1
R7	P16∧P17∧P8∧P7	0.316	P1
R8	P16∧P17∧P8∧P3	0.316	P1
R9	P16∧P17∧P8∧P4	0.316	P1

4.2 *Rule antecedent analysis*

Pattern discovery from rule antecedents was initiated from three-sequence itemsets until six-sequence itemsets since there was a steady decline in the number of generated rules with high sequence itemsets. We believe this indicated that the frequency of visits to those was negligible. Table 3 indicates the rules generated as three-sequence itemsets based on the highest lift of each rule. The lift enables predicting the probable hits of the user during the subsequent visits to the website. This relates not to the order of access but the most possible sequence of visit.

The single occurrence rules for three sequence itemsets were ignored since the frequency of the visit is negligible. The rule R1 (P16∧P17∧P13) with the minimum support of 0.0000152 had occurred 11 times with different pages, in which the highest lift (0.316402) focuses on the page P1, the rule R2 (P16∧P17∧P15) occurred five times in which the highest lift (0.219335) focuses on page P6 as the next access page, the rule R3 (P16∧P17∧P8) occurred for six times with different consequents and page P1 was viewed as the subsequent page with the highest lift (0.316402). Similarly the rule R4 (P16∧P17∧P8) occurred 11 times, also denotes the page P1 as

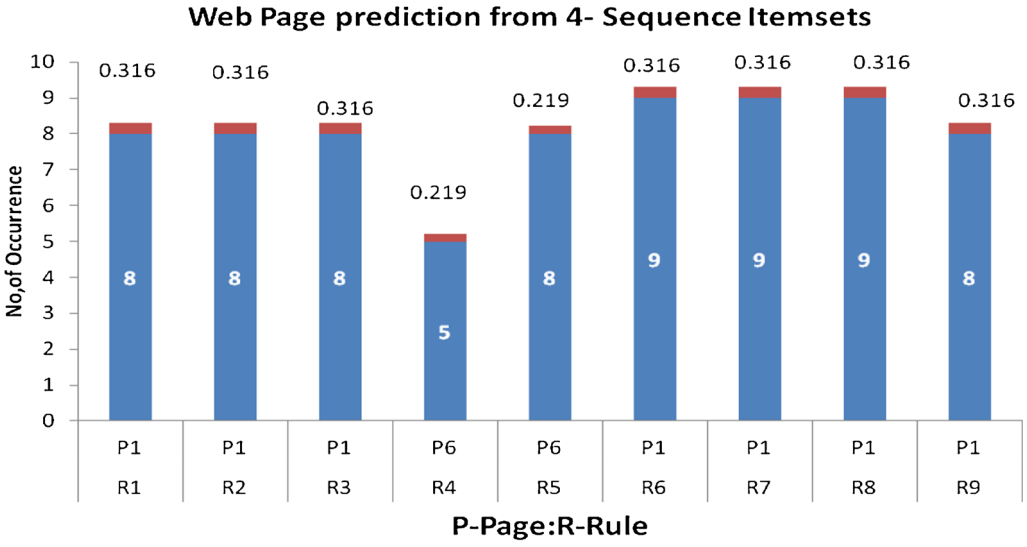

Figure 3. Webpage prediction from 4-sequence itemsets.

the subsequent page with the highest lift 0.316402. In all these case we filtered the consequent page based on the interesting measure 'lift'. From the rules R1, R3 and R4 it is clearly identified that the user with the navigation pattern from MSN-News (P16) followed by MSN-Sports (P17) and then with Summary (P13) or Wealth (P8) or Health (P9) will subsequently access the Front-Page (P1). The rule R2 states that the user with the pattern from MSN-News and Sports with travel will surely access the On-air (P6) page as the subsequent page.

The result based on three sequence itemsets reveals that the pages P1 and P6 are considered as more promising next access page when compared to other pages. The graphical representation is depicted in figure 2.

There are 922 rules generated in four sequence itemsets. Discovering pattern from vast rules is a tedious process, so we filtered the rules based on the minimum support threshold. The importance of the rules generated for four-sequence itemsets with the minimum support of 0.0000495 based on the highest lift is tabulated in table 4.

The results clearly predict that the user who visited MSN-News will visit MSN-Sports and then Travel and so on with the pages P1 (Front page) and P6 (O-Air) as the probable subsequent access pages.

The graphical representation of the itemsets from four-sequence page visits is depicted in figure 3.

Table 5. Pattern discovery from rule antecedents in 5-sequence itemsets.

Rule no.	Rule	Highest lift	Predicted subsequent page
R1	P17∧P5∧P13∧P10∧P12	0.219	P6
R2	P17∧P5∧P13∧P10∧P4	0.219	P6
R3	P17∧P5∧P13∧P10∧P6	0.177	P2
R4	P17∧P5∧P13∧P10∧P2	0.219	P6
R5	P17∧P5∧P13∧P11∧P12	0.219	P6
R6	P17∧P5∧P13∧P6∧P1	0.177	P2
R7	P17∧P5∧P13∧P9∧P12	0.219	P6
R8	P17∧P5∧P13∧P12∧P1	0.177	P2
R9	P17∧P15∧P10∧P8∧P9	0.316	P1
R10	P17∧P15∧P10∧P9∧P14	0.316	P1
R11	P17∧P15∧P8∧P11∧P9	0.316	P1
R12	P17∧P15∧P8∧P9∧P14	0.316	P1
R13	P17∧P15∧P8∧P9∧P3	0.316	P1
R14	P17∧P13∧P10∧P11∧P9	0.316	P1
R15	P17∧P13∧P10∧P8∧P7	0.219	P6
R16	P17∧P13∧P10∧P7∧P6	0.12	P14
R17	P17∧P13∧P10∧P7∧P2	0.219	P6
R18	P17∧P13∧P10∧P11∧P7	0.123	P4
R19	P17∧P13∧P10∧P7∧P12	0.123	P4
R20	P17∧P13∧P10∧P7∧P4	0.12	P14
R21	P17∧P13∧P10∧P11∧P3	0.12	P14
R22	P17∧P13∧P8∧P7∧P12	0.219	P6
R23	P17∧P13∧P8∧P12∧P3	0.219	P6
R24	P17∧P13∧P8∧P3∧P6	0.177	P2
R25	P17∧P13∧P8∧P3∧P2	0.219	P6
R26	P17∧P13∧P7∧P3∧P2	0.219	P6
R27	P17∧P13∧P9∧P7∧P3	0.316	P1

The numbers of rules generated by five sequence itemsets are 7403. Table 5 indicates the importance of the rules generated for five-sequence itemsets based on the lift with the minimum support of 0.0000192. Rule R1 (P17∧P5∧P13∧P10∧P12) appears three times with different consequent pages whereas the page P6 is considered as the subsequent page with the highest lift of 0.219.

The result revealed that web pages P6, P1, P2, P4 and P14 are the most interesting subsequent pages based on the navigation behaviour of the user interest. It is needful to note that the process of determining the subsequent probable page of visit by a web user is similar to the 3-sequence, 4-sequence and 5-sequence patterns. This reveals the fact that the proposed methodology is applicable to predict the subsequent pages for sequence of any length.

The graphical representation of the itemsets from five-sequence page visits is depicted in figure 4.

Similarly the six sequence itemset rules were generated based on highest lift and the number of occurrence of rules. The result shows that P1, P6, P14 and P12 are highly interesting pages from the navigation behaviour of the user interest. The above pattern discovery from rule antecedents clearly shows that pages P1 and P6 are highly interesting subsequent pages and P1, P2, P4, P6, P12, P14 fall under one category.

4.3 *Rule consequent analysis*

Page visit patterns analyzed from Rule consequents was done based on the webpage that served as the consequent for the rule. Totally 467,384 rules were generated with the confidence of 100% and the support of 1.3E−05. Since the number of rules generated was vast, we filtered all the rules that were considered for analysis satisfying the minimum support of 0.0000131 and

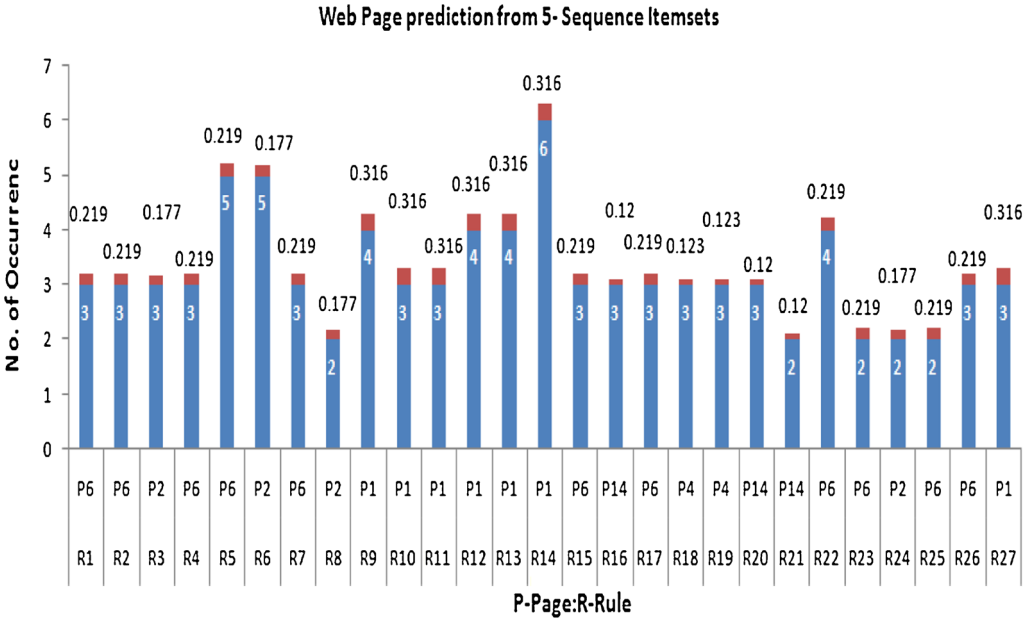

Figure 4. Webpage prediction from 5-sequence itemsets.

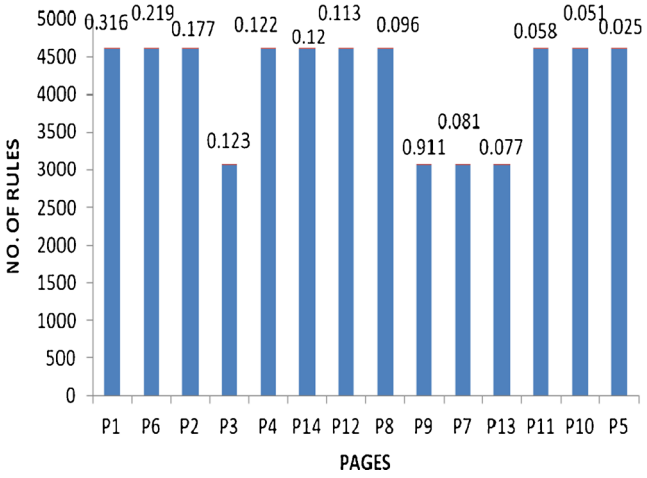

Figure 5. Pages and no. of rules based on rule consequents.

confidence of 100%. Each website page was analyzed individually to detect the frequency of visits for that particular page.

Based on the aforementioned filtering criterion, the number of rules pertaining to each website page visited was determined. Figure 5 portrays the graphical representation of significance of the pages based on the number of rules that targeted the particular page visits. The pages that were detected to be highly interesting to the users appeared to be of two categories. Pages that generated rule sets containing 4608 rules and the lesser interesting pages containing 3072 rules. Pages P1, P2, P4, P5, P6, P8, P10, P11, P12, P14 fell under the first category whereas pages P3, P7, P9 and P13 were placed under the second category. Also revealed was the fact that P15, P16 and P17 were not subsequently being accessed by any rule with the access pattern containing the pages included in any of the above categories.

From the analysis on the patterns of rules generated by Apriori-PT, it was evident that the utilization of the association rule mining methodology for web usage pattern mining revealed several facts on the trends of web usage by different users throwing light on the web browser's areas of interest.

5. Conclusion

Utilization of data mining methodologies in prediction of web usage mining (Babu 2000) has been an area of intense research in recent years. This research was undertaken to predict patterns of user's visit to web pages that were listed in the MSNBC dataset. The dataset was voluminous and hence the number of rules generated was vast. Apriori-PT algorithm was utilized to mine the frequent pattern rules from which patterns were analyzed based on the rule consequents and antecedents. Voluminous rules were filtered based on the minimum support threshold. Moreover pattern discovery was based on the number of occurrences of the rule and the corresponding web page visited. This research revealed with 100% confidence and with the high value of Lift that P1, P2, P4, P5, P6, P8, P10, P11, P12, P14 were the pages that were more subsequently accessed by the users when compared with the pages P3, P7, P9 and P13. Pages P15, P16 and P17 were not targeted as the subsequent access page by any interesting rules.

References

Agrawal R and Srikant R 1994 Fast algorithms for mining association rules. In: *Proceedings of 20th International Conference on Very Large Data Bases*, VLDB, 1215(1): 487–499

Anitha A and Krishnan N 2011 A dynamic web mining framework for E-learning recommendations using rough sets and association rule mining. *Int. J. Comp. Appl.* 12(11): 19–25

Babu K G, Komali A, Mythry V and Ratnam A S K 2000 Web mining using semantic data mining techniques. Int. *J. Soft Comput. Eng. (IJSCE)* 3(2): 2231–2307

Chakrabarti S 2002 Mining the web: Analysis of hypertext and semi structured data, Morgan Kaufmann

Chandra B and Basker S 2000 A new approach for classification of patterns having categorical attributes. *IEEE International Conference on Systems, Man, and Cybernetics (SMC)* 960–964

Chifu V and Salomie I 2009 A fluent calculus approach to automatic web service composition. *Adv. Electr. Comput. Eng.* 9(3): 75–83

Chun-sheng Z and Li Yan 2014 Extension of local association rules mining algorithm based on Apriori algorithm, pp. 340–343

Debahuti M 2010 Predictive data mining: Promising future and applications. *Int. J. Comput. Commun. Technol.* 2(1): 20–28

Eirinaki M, Vazirgiannis M and Kapogiannis D 2005 Web path recommendations based on page ranking and Markov models. In: *Proceedings of the 7th Annual ACM International Workshop on Web Information and Data Management*, 2–9

Ganapathy S, Sethukkarasi R, Yogesh P, Vijayakumar R and Kannan A 2014 An intelligent temporal pattern classification system using fuzzy temporal rules and particle swarm optimization. *Sadhana, Indian Acad. Sci.* 39(2): 283–302

Gao S, Alhajj R, Rokne J and Guan J 2009 Set-based approach in mining sequential patterns. In: IEEE *24th International Symposium on Computer and Information Sciences, 2009. ISCIS 2009.* pp. 218–223

Hacibeyoglu M, Arslan S and Kahramanli S 2013 A hybrid method for fast finding the reduct with the best classification accuracy. *Adv. Electr. Comput. Eng.* 13(4): 57–64

Han J and Kamber M 2011 *Data mining – Concepts and techniques*, 3rd edition, Morgan Kauffmann Publishers

Hung Y S, Chen K L B, Yang C T and Deng G F 2013 Web usage mining for analysing elder self-care behavior patterns. *Expert Syst. Appl.* 40(2): 775–783

Kum Hye-Chung, Paulsen Susan and Wang Wei 2005 Comparative study of sequential pattern mining frameworks -support framework vs. multiple alignment framework. In IEEE *2nd International conference on data mining* - workshop on the *foundation of data mining and discovery*. *ICDM 2002*. pp. 43–70

Internet Usage Statistics http://www.internetworldstats.com/stats.htm

Jacob S G and Ramani R G 2012 Evolving efficient classification rules from cardiotocography data through data mining methods and techniques. *Eur. J. Sci. Res.* 78(3): 468–480

Jacob S G and Ramani R G 2013 Design and Implementation of a clinical data classifier: A supervised learning approach. *Res. J. Biotechnol.* 8(2): 16–24

Jacob S G, Ramani R G and Nancy P 2013 Discovery of knowledge patterns in lymphographic clinical data through data mining methods and techniques. *Advances in computing and information technology.* LNCS Springer Berlin Heidelberg, 129–140

Jaideep S, Cooley R, Deshpande M and Tan P N 2000 Web usage mining: Discovery and applications of usage patterns from web data. *ACM SIGKDD Explorations Newsletter* 1(2): 12–23

Kotsiantis S B and Kanellopoulos D 2001 Association rules mining: A recent overview. *GESTS Int. Trans. Comput. Sci. Eng.* 32(1): 71–82

Kotsiantis S B, Zaharakis I D and Pintelas P E 2007 Supervised machine learning: A review of classification techniques, pp. 3–24

Kriegel H P 2007 Future trends in data mining. *Data Mining Knowledge Discovery* 15(1): 87–97

Kumar S K and Chezian R M 2012 A survey on association rule mining using Apriori algorithm. *Int. J. Comput. Appl.* 45(5): 7–50

Liu L and Peng T 2013 Post-processing of deep web information extraction based on domain ontology. *Adv. Electr. Comput. Eng.* 13(4): 25–32

Madhuri B 2002 Analysis of the navigation behavior of the users' using grey relational pattern: Analysis with Markov chains. *Int. J. Eng. Sci. Technol.* 2(10): 5402–5412

Mary S S A and Malarvizhi M 2012 A new improved weighted association rule mining with dynamic programming approach for predicting a user's next access. *Comput. Sci. Inform. Technol.* 2(1): 10–15

Mitchell T 2009 *Machine learning*. McGraw Hill

Phoa F K H and Sanchez J 2013 Modeling the browsing behavior of world wide web users. *Open Journal of Statistics*. 3(2):145–154

Ramani R G and Jacob S G 2013a Improved classification of lung cancer tumors based on structural and physicochemical properties of proteins using data mining models. *PloS one* 8(3): e58772

Ramani R G and Jacob S G 2013b Benchmarking classification models for cancer prediction from gene expression data: A novel approach and new findings. *Studies Informatics Control* 22(2): 134–143

Ramani R G, Lakshmi B and Jacob S G 2012 Data mining method of evaluating classifier prediction accuracy in retinal data. *IEEE International Conference on Computational Intelligence & Computing Research (ICCIC)*

Renáta I and Vajk I 2006 Frequent pattern mining in web log data. *Acta Polytechnica Hungarica* 3(1): 77–90

Robert C, Mobasher B and Srivastava J 1999 Data preparation for mining world wide web browsing patterns. *Knowledge Inform. Syst.* 1(1): 5–32

Sanchez J and Liu C T 2011 Bayesian hierarchical model of the browsing behavior of world wide web Users. Department of Statistics, UCLA

Santhisree K and Damodaram A 2010 Optics on sequential data: Experiments and test results. *Int. J. Comput. Appl.* 11(5): 15–21

Suraya A, Norhisham R M and Fun T S 2011 Discovering frequent sequential pattern using personalized minimum support threshold with minimum items. *International Conference on Research and Innovation in Information Systems (ICRIIS)* 10(1): 1–6

Suresh K, Madanamohana R, Reddy R A and Subramanyam A 2011 Improved FCM algorithm for clustering on web usage mining. *IEEE International Conference in Computer And Management (CAMAN)* 1–4

Tassa T 2014 Secure mining of association rules in horizontally distributed databases. *IEEE Trans. Knowledge Data Eng.* 26(4): 970–983

University of California, Machine Learning Repository https://archive.ics.uci.edu/ml/.../MSNBC.com+Anonymous+Web+Data

Veeramalai S, Jaisankar N and Kannan A 2010 Efficient web log mining using enhanced Apriori algorithm with hash tree and fuzzy. *Int. J. Comput. Sci. Inform. Technol.* 2(4): 241–247

Wang W, Yang J and Philip S Y 2000 Efficient mining of weighted association rules (WAR) In: *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 13–31

Wen-Hai G 2010 Research on client behavior pattern recognition system based on web log mining. *International Conference On Machine Learning and Cybernetics (ICMLC)* 1(1): 10–21

Yang B, Xiangjun D and Fufu S 2009 Research of web usage mining based on negative association rules. *International Forum on Computer Science-Technology and Applications* 1(1): 336

Zhang Y and Chen G 2014 A Forensics method of web browsing behaviour based on association rule mining. In: *2nd International Conference on Systems and Informatics*, pp. 927–932

Zhou X and Huang Y 2014 An improved parallel association rules algorithm based on mapreduce framework for big data. In: *11th International Conference on Fuzzy Systems and Knowledge Discovery*, pp. 284–288