Mining Association Rules from Web Logs by Incorporating Structural Knowledge of Website

Bhawna Nigam
Department of Information Tech.
IET- DAVV, Indore, India

Suresh Jain
Department of Computer Engg.
KCB Technical Academy, Indore, India

Sanjiv Tokekar
Department of Electronics & Telecommunication Engineering
IET- DAVV, Indore, India

ABSTRACT

One of the basic problems with the Association Rule discovery is that when Mining Algorithms are applied on Web Access Logs, the total number of generated rules is found to be very large. For finding useful results from these rules, the analyzer needs to look into large rule-set. Moreover, the analysis of such rule-set also requires certain criteria for making decisions, i.e. a particular rule should be accepted or not. This ambiguity of acceptance or rejection of rules makes it very difficult to extract knowledge. Hence in order to get effective results with the minimized effort, number of rules should be less and valid. Therefore, the structural knowledge of Website is considered to solve the purpose, that plays an important role in pruning the invalid rules, thereby reducing the size of rule-set , and it is observed from the experiment that the number of rules have been successfully reduced.

Keywords

Association Rules, Weblog, Web Usage Mining, Website Structure, Trails or Navigation Session, BFS(Breadth First Search)

1. INTRODUCTION

Web Usage Mining is one of the categories of Web Mining that helps the Analysts, Web Masters and Users to find the navigational patterns in order to understand the needs of Web based Applications in a better way [1] [2]. Web Usage Mining plays a very important role for the Software Organizations in maintaining the E-Business Applications, in performing Personalization of Website and also helps Web Masters to analyze, manage and modify the site according to the current needs. The processed Web Usage Data also helps the Business Organizations in analyzing the previous patterns of expenditure and sales so that they can produce more effective results to their business and increase their future sales. Usage data can also be useful for developing strategies according to the current market trends that will promote the Company's services or product on a higher level to sustain in a competitive environment [15]. It also helps in managing effective relationship with the Customers[16].Various researches are been performed in this field, such as modifying and improving a Web Site's Design by finding unexpected browsing behavior in Click Stream data [3], improving overall effectiveness of Web Based Teaching and Learning [4], Business Intelligence, Path Completion[17] etc.

Web Usage log files generated on the Server can be used for finding Association Rules, which can give information to the Data Analysts such as the interest of users, the failure of page requests, the amount of time that the user spends on the site etc. Some of the applications of Association Rule in Web Usage Mining are E-Commerce Applications , in which the frequent Access patterns of User's Visit can be seen and analyzed for Site Improvement [11] [12] [14], proposal of Recommender System for personalization in adaptive Web-Based Applications[13], Prediction of Webpage Accesses[8] etc.

For finding the Association Rules in context of Web Usage Mining, Researchers have used various techniques, the most common of which are Apriori Algorithm and Frequent Pattern Growth Algorithm [2][5]. Mei-Ling Shyu incorporated the concept of Minimum Reaching Distance for pruning rules [6] in Apriori Algorithm. Yuhua Chen, Xin Chen and Haoyi Chen had introduced the Maximal Forward Reference in Apriori Algorithm, and used the precision, coverage and F1 measure for measuring recommendation effectiveness [9]. D. Vasumathi & Dr. A Govardhan used the concept of ordered lattice to mine the Association Rules [7]. Faten Khalil, Jiuyong Li and Hua Wang proposed the concept of Markov Model to find the Association Rules for predicting Web Page access [8]. Jose Borges also used the Markov Model, along with the concept of Hypertext Probabilistic Grammar, but has generated the rules using the graph exploration method i.e. Breadth First Search Approach [10].

The problem with the Association Rule Mining is that the count of generated rules obtained from Web Logs is very large. Also, the validity of the generated rules is a big question. The reasons for not getting correct Association Rules are –

1. The log files obtained may be incomplete: The completeness of Web Log is always undecidable when cache entries came into picture. For e.g. if referrer based method is used to find the navigational session, and suppose the same web page is requested by same user using the same referrer web page, then such entry will not be recorded in the log file. Hence incomplete logs will be obtained in that case. Moreover all the fields are not applicable to every user accessing the web site. For e.g. when a user simply visits the website, then User ID may not be required. So the inapplicable fields are left blank in the log entry. Due to such incompleteness of field entries, assumptions are to be made to identify user or navigation sessions.

2. The navigation sessions obtained from log files may be

incorrect as the known pre-processing methods may or may not generate correct trails: The identification of user and session can be done using various methods, based on the fields of access logs, such as IP Address, Agent field or Cookie field or combination of them. But each method may or may not get fitted into various situations and thus results in incorrect trail information. For e.g. IP address is a strong parameter to identify user. But when “N” number of computers connected in a LAN, they own same gateway IP address. When various users visit the site using the same LAN connection, using the same browser, then it becomes very difficult to distinguish the users using this parameter only. Or it may happen that the cookies are disabled at the user’s end. So the field will be left blank in the log entry. Therefore it is not guaranteed that the trails obtained are always correct.

So, Validation of the obtained rules is done on the basis of the topology of the Website. The topology of a website is the structure of website, stored in the form of a matrix that shows the connectivity among web pages , i.e. whether there is path between 2 web pages or not.

2. METHODOLOGY

Three different approaches are analyzed in this paper, based on the number of rules mined. In the first approach i.e. BFS approach, the Hypertext Transition Probability Matrix is prepared from the Web Log and the BFS algorithm is applied to find the Association Rules. Other methods use the structural knowledge of the Website for mining the Association Rules from Weblog. Some important concepts-

Trail or User Navigation Session

Trail is the sequence of web pages actually visited by the user. This information is obtained from Web Log after Preprocessing. For e.g. a1 a2 a3 is a trail, which means that User visited web pages in the order a1, a2 and a3.

Hypertext Transition Probability

It refers to the probability that one Web Page is visited after another. For e.g. when user visited Page a2 after Page a1 in a website (i.e. a1->a2) then Hypertext Transition Probability = (no. of times a2 is visited after a1)/ (no. of times a1 is visited)

Hypertext Transition Probability Matrix

The values of Hypertext Transition Probabilities are stored in the form of matrix for further calculation of Association Rules.

Derivation Probability

Derivation Probability means the product of probabilities of productions used in deriving the Association Rules.

Cut Off Point

It is the minimum value that should possess by the derivation probability of a rule. A rule is considered if its derivation probability is above a given cut off point. Its value can be derived as - Cut off Point = Support * Confidence

Website Topology

The structure of a sample website is shown in Fig 1. There are three Web Pages present in the Website i.e. a1, a2 and a3. “start” and “final” are conceptual states from where, user’s visit to the site is assumed to be begin and end.

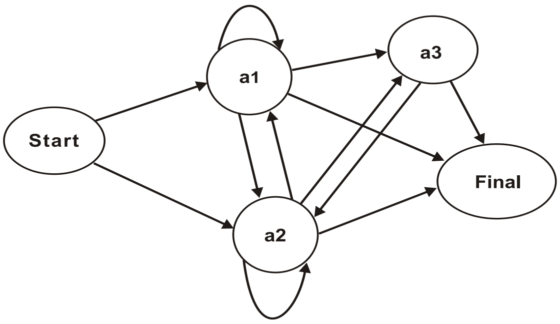

Fig. 1: Structure of Website

Table 1 shows the Website Structure Matrix prepared from the Website topology. In Matrix, ‘0’ shows the absence of the hyperlink and ‘1’ shows the presence of the hyperlink between the WebPages.

Table 1. Website Structure Matrix

	start	a1	a2	a3	final
start	0	1	1	0	0
a1	0	1	1	1	1
a2	0	1	1	1	1
a3	0	0	1	0	1
final	0	0	0	0	0

Methodology I: BFS approach

Till now, many Researchers have used the BFS approach to find Association Rules. The steps are-

1. Preprocess the Web Log (includes data cleaning, user identification and session identification) and obtain the trail information.
2. Obtain the Hypertext Transition Probability Matrix from trail information.
3. Generate the Association Rules using BFS approach by traversing the Hypertext Transition Probability Matrix, starting from “start” state till the derivation probability goes below the cut-off point.

Example 1: Trails obtained from web log (for the website shown in Fig.1) after preprocessing are-

a1 a2 a3 (start->a1->a2->a3->final)
a2 a3 (start->a2->a3->final)
a3 a2 (start->a3->a2->final)

The above example shows three trails traversed by the user. Table 2 shows the corresponding Hypertext Transition Probability Matrix.

Table 2. Hypertext Transition Probability Matrix for Example 1

	start	a1	a2	a3	final
start	0	1/3 = 0.33	1/3 = 0.33	1/3=0.33	0
a1	0	0	1/1=1	0	0
a2	0	0	0	2/3 = 0.66	1/3 = 0.33
a3	0	0	1/3 = 0.33	0	2/3=0.66
final	0	0	0	0	0

Cut off Point = Support * Confidence

For finding the Association rules from the above matrix, BFS method is used. The exploration begins from "start" state and will continue until the derivation probability falls below the cut-off point. Let the cut-off point for sample data be 0.2. Fig 2 shows the BFS exploration tree for the Hypertext Transition Probability Matrix shown in table 2. The "X" in diagram means that the derivation probability goes below the cut-off point hence that branch will not be explored further.

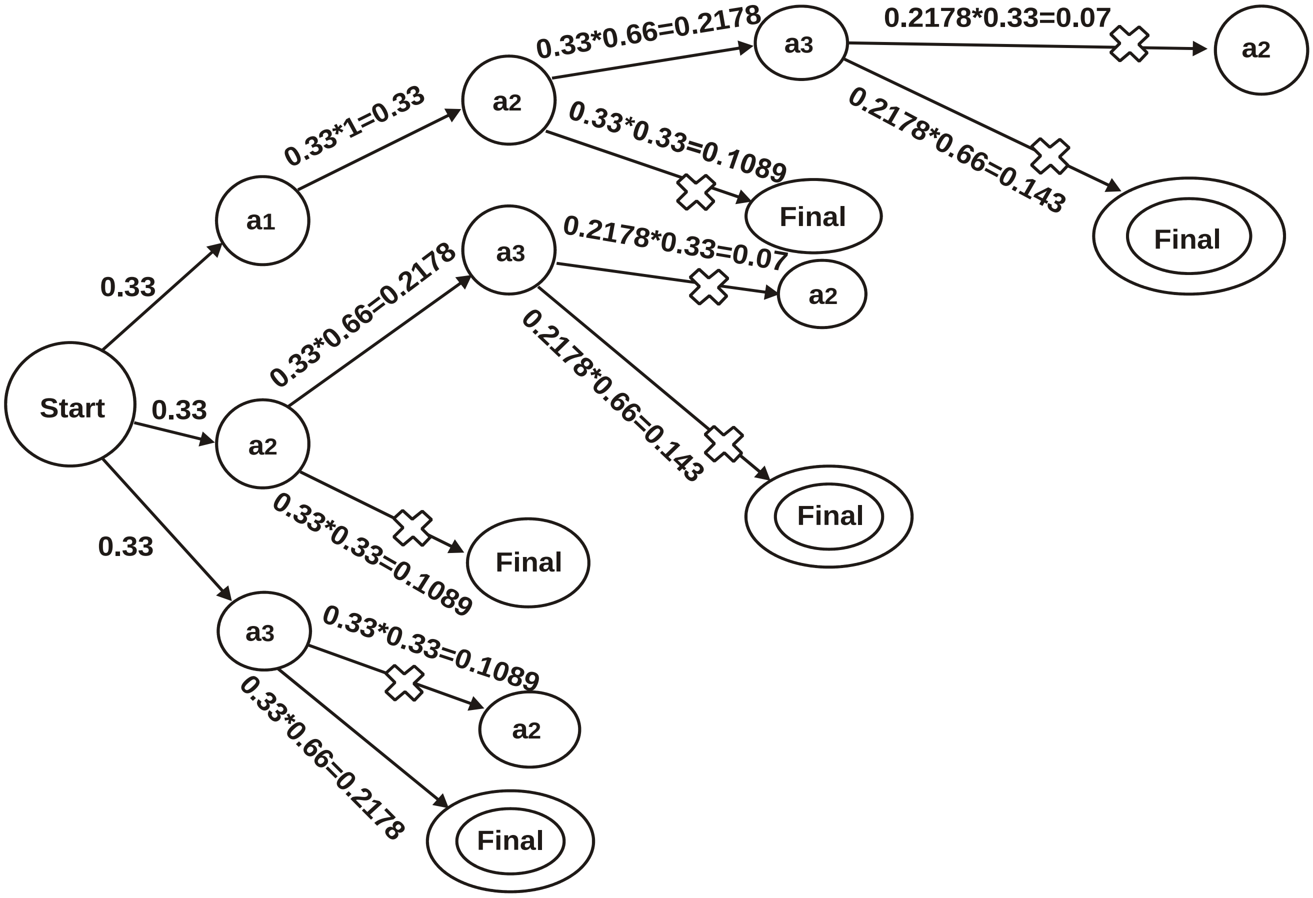

Fig. 2. BFS exploration tree for Example 1 (Table 2) having Cut-off point=0.2.

Result: Generated Association rules are

1. start->a3->final = 0.2178
2. start->a2->a3= 0.2178
3. start->a1->a2->a3= 0.2178

Methodology 2: Discarding Invalid trail (DIT) before Rule Mining

In this approach the knowledge about the Website Structure is used to discard the invalid trails. The steps for mining the Association Rules are -

1. Preprocessing of Web Log (includes data cleaning, user identification and session identification). The results obtained are user trails.

2. Create the topology matrix from the structure of website.

3. Read the trail information and discard Invalid trails according to the topology (if the path is not available for the trail), before Association Rule Mining.

4. Create Hypertext Transition Probability Matrix using the trail information.

5. Apply BFS Algorithm to get the sequential rules, starting from "start" state until the derivation probability falls below the cut-off point.

Considering the trails mentioned in Example 1, the trails are-

a1 a2 a3 (start->a1->a2->a3->final)
a2 a3 (start->a2->a3->final)
a3 a2 (start->a3->a2->final)

Since start->a3->a2->final is invalid trail because there is no path exists between start to a3, as shown in the topology (Fig.1.) hence it is discarded. Now with two valid trails shown below the transition probability matrix is prepared.

a1 a2 a3 (start->a1->a2->a3->final)
a2 a3 (start->a2->a3->final)

Table 3. Hypertext Transition Probability Matrix for above trail information

	start	a1	a2	a3	final
start	0	1/2 = 0.5	1/2 = 0.5	0	0
a1	0	0	1/1=1	0	0
a2	0	0	0	2/2 = 1	0
a3	0	0	0	0	2/2= 1
Final	0	0	0	0	0

Again considering cut-off value equal to 0.2, Fig 3 shows BFS exploration tree. Because all the generated rules have the derivation probability above the cut off point so they are considered as valid. Result: Generated association rules are

1. start->a1->a2->a3->final = 0.5
2. start->a2->a3->final= 0.5

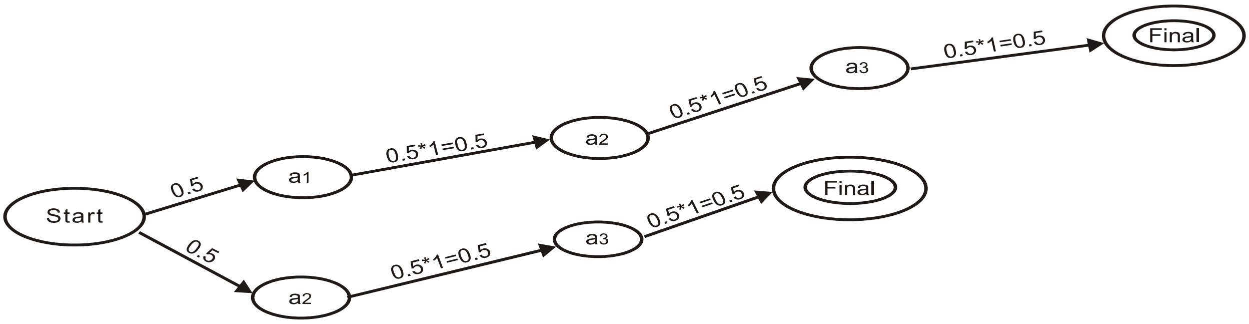

Fig. 3: BFS exploration tree for the same example using Methodology 2.

Methodology 3: Discarding Invalid Association Rule (DIAR)

All the valid or invalid trails are considered to find the rules. If the path for a particular rule, generated in the rule-set is not available in topology of the site, the Association rule will not be considered further. The steps are-

1. Preprocessing of Web Log (includes data cleaning, user identification and session identification). The results obtained are user trails.
2. Create the website topology matrix from the structure of the website.
3. Prepare Hypertext Transition Probability Matrix from the obtained trail information.
4. Apply BFS Algorithm to obtain the Association Rule until the derivation probability is less than the cut-off point.
5. Match each of the obtained Association Rule with the topology of the website. If there is no such path exists, discard the association rule.

Again considering Example 1 i.e.

a1 a2 a3 (start->a1->a2->a3->final)
a2 a3 (start->a2->a3->final)
a3 a2 (start->a3->a2->final)

The Hypertext Transition Probability Matrix is generated as shown in Table 2. Fig 1 shows BFS exploration tree for Example 1 for cutoff point 0.2. Generated Association Rules are-

start->a3->final= 0.2178
start->a2->a3= 0.2178
start->a1->a2->a3= 0.2178

If the generated association rule is not found in the topology, then the association rule will be pruned. In above example, there is no path for start->a3->final, hence it will be discarded.

So, final generated rules are

start->a2->a3= 0.2178
start->a1->a2->a3= 0.2178

Table 4 shows Association Rules generated by BFS Approach, Discarding Invalid Trail and Discarding Invalid Association Rule (DIAR). With the incorporation of the structural knowledge of Website, minimum numbers of valid Association rules are generated.

Table 4. Association Rules generated by (i) BFS Approach (ii) Discarding Invalid Trail (DIT) (iii) Discarding Invalid Association Rule (DIAR)

	BFS Approach	**Discarding Invalid Trail**	**Discarding Invalid Association Rule (DIAR)**
Generated Association Rules	1.start->a3->final = 0.2178 2.start->a2->a3= 0.2178 3.start->a1->a2->a3= 0.2178	1.start->a1->a2->a3->final=0.5 2.start->a2->a3->final= 0.5	1. start->a2->a3= 0.2178 2.start->a1->a2->a3= 0.2178

3. EXPERIMENTAL RESULT

Log Files: The log files are taken from a Software firm, WebSkyInfotech, Indore. The server access logs of 21 days (for the month of October 2011) are analyzed for the results. The site is designed in ASP.NET, and the log files are in the W3C Extended Log Format (IIS 6.0). The User Navigation Sessions or Trail Information is generated as a result of Pre-processing of the log files. To get the trail information, the individual log files in .txt format are analyzed first, using the Java Code. So the number of trails obtained after pre-processing are-

Log file of 1-oct-2011-> 36 trails
Log file of 2-oct-2011-> 53 trails
Log file of 3-oct-2011-> 20 trails
Log file of 4-oct-2011-> 24 trails

```
#Software: Microsoft Internet Information Services 6.0
#Version: 1.0
#Date: 2011-10-01 04:26:12
#Fields: date time s-sitename s-computername s-ip cs-method
cs-uri-stem cs-uri-query s-port cs-username c-ip cs-version
cs(User-Agent) cs(Cookie) cs(Referer) cs-host sc-status sc-
substatus sc-win32-status sc-bytes cs-bytes time-taken
2011-10-01 04:26:12 W3SVC461857616 SERVER9X
174.37.211.193 GET /wsi/page/ourstrenght.aspx – 80 -
180.76.5.141 HTTP/1.1
Mozilla/5.0+(compatible;+Baiduspider/2.0;++http://www.bai
du.com/search/spider.html) - - webskyinfotech.com 200 0 64
2048 245 625
```

Example of Web Log used during experiment

Log file of 5-oct-2011-> 13 trails and so on....

In order to analyze the results in a uniform format, log files are clubbed together, to get the number of trails as 100, 200, 300, 400 and 500.

TOPOLOGY

The topology is stored in a text file by traversing the website. There are 33 pages in the Website. It is assumed that the topology remains the same for the month of October.

The log files are analyzed for 5 different cut off values. The following table shows the number of generated rules for Approach – I i.e. Normal Approach, in which rules are evaluated but not validated, Approach II- Discarding Invalid Trails before Association Rule Mining and Approach III- Discarding Invalid Association Rules after mining. For different cut-off values, the comparative analysis of all the approaches is shown in Fig. 4

I. BFS APPROACH or NORMAL APPROACH (NA)

NUMBER OF TRAILS	**NUMBER OF RULES GENERATED FOR FOLLOWING CUT OFF POINT VALUES**				
	0.1	**0.01**	**0.001**	**0.0001**	**0.00001**
100	1	23	73	138	289
200	1	24	63	108	201
300	1	25	74	161	357
400	1	28	92	249	762
500	1	29	94	277	827

II. DISCARDING INVALID TRAILS (DIT)

NUMBER OF TRAILS	**NUMBER OF RULES GENERATED FOR FOLLOWING CUT OFF POINT VALUES**				
	0.1	**0.01**	**0.001**	**0.0001**	**0.00001**
100	1	17	33	63	125
200	1	12	23	55	87
300	1	12	39	96	286
400	1	16	52	185	666
500	1	16	47	182	627

III. DISCARDING INVALID ASSOCIATION RULES (DIAR)

NUMBER OF TRAILS	**NUMBER OF RULES GENERATED FOR FOLLOWING CUT OFF POINT VALUES**				
	0.1	**0.01**	**0.001**	**0.0001**	**0.00001**
100	1	9	31	72	160
200	1	6	27	50	104
300	1	6	30	75	182
400	1	9	38	112	355
500	1	9	37	100	337

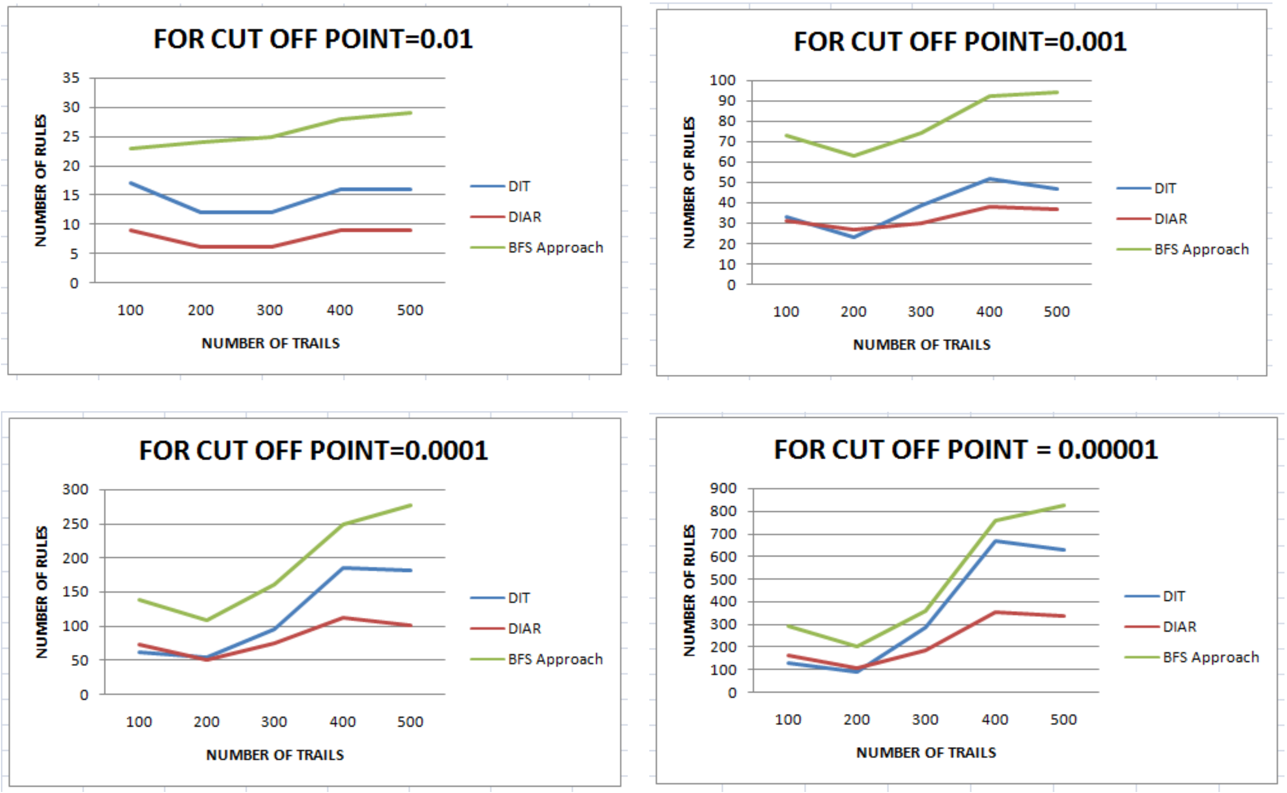

Fig.4. Number of rules generated by i) DIT ii) DIAR iii) BFS Approach on different cut off points.

CONCLUSION

From the experimental results shown in the tables and the Figure 4, it can be observed that using topology of website, the number of rules has been successfully reduced in the proposed approaches DIT and DIAR as compared to the Normal BFS Approach. It can also be noted that as the value of Cut-Off point decreases, the number of rules are going to be increased in all the approaches. Comparing the approaches on the basis of total number of rules generated, we can see that the DIAR approach results in less number of rules as compared to that of Normal BFS Approach and DIT approach. So DIAR is the most appropriate method for obtaining minimized rule-set. The future work regarding this paper is to use the Statistical approaches for exhaustive analysis of generated rules. Also, we will try to analyze the discarded rules in order to find whether they are of certain importance or not.

REFERENCES

[1] Jaideep Srivastava , Robert Cooleyz , Mukund Deshpande, Pang-Ning Tan : *Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data*. ACM SIGKDD (2000).

[2] B.Santhosh Kumar, K.V.Rukmani : *Implementation of Web Usage Mining Using APRIORI and FP Growth Algorithms.* Int. J. of Advanced Networking and Applications 400 Volume: 01, Issue: 06, Pages: 400-404 (2010)

[3] I-Hsien Ting, Chris Kimble, Daniel Kudenko: *UBB Mining: Finding Unexpected Browsing Behaviour in Clickstream Data to Improve a Web Site's Design.*

[4] A.Anitha, N.Krishnan: *A Web Usage Mining based Recommendation Model for Learning Management Systems.*CONFERENCE- IEEE (2010)

[5] Huiping Peng:*Discovery of Interesting Association Rules Based on Web Usage Mining* International Conference on Multimedia Communications (2010)

[6] Mei-Ling Shyu, Choochart Haruechaiyasak, Shu-Ching Chen and Na Zhao: *Collaborative Filtering by Mining Association Rules from User Access Sequences*- IEEE (2005)

[7] D. Vasumathi and Dr. A Govardhan: *Efficient Web Usage Mining Based on Formal Concept Analysis.* Journal of Theoretical and Applied Information Technology (2005 – 2009)

[8] Faten Khalil, Jiuyong Li and Hua Wang: *A Framework of Combining Markov Model with Association Rules for Predicting Web Page Accesses*. Proc. Fifth Australasian Data Mining Conference (2006)

[9] Yuhua Chen, Xin Chen and Haoyi Chen: *Improve on Frequent Access Path Algorithm in Web Page Personalized Recommendation Model.* International Conference on Information Science and Technology Nanjing, .Jiangsu, China (March 26-28, 2011)

[10] Jos´e Lu´ıs Cabral de Moura Borges: *A Data Mining Model to Capture User Web Navigation Patterns.* (2000)

[11] Shaofei Wu: *A New Frequent Path Algorithm of Web*

*User Access Pattern.*International Conference on Industrial and Information Systems (2009)

[12] S. Taherizadeh N. Moghadam : *Integrating Web Content Mining into Web Usage Mining for Finding Patterns and Predicting Users' Behavior.* International Journal of Information Science and Management

[13] Daniel Mican, Nicolae Tomai : *Association-Rules-Based Recommender System for Personalization in Adaptive Web-Based Applications.*

[14] Maja Dimitrijević, Zita Bošnjak : *Web Usage Association Rule Mining System.*Interdisciplinary Journal of Information, Knowledge, and Management Volume 6 (2011)

[15] Web Data Mining. Net http://www.web-datamining.net/usage/

[16] Olfa Nasraoui, Esin Saka, Antonio Badia and Richard Germain: *A Web Usage Mining Framework for MiningEvolving User Profiles in Dynamic Web Sites.*IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 20, NO. 2 (2008)

[17] Yan LI, Boqin FENG, Qinjiao MAO: *Research on Path Completion Technique inWeb Usage Mining.* International Symposium on Computer Science and Computational Technology (2008).