IJournals: International Journal of Software & Hardware Research in Engineering
ISSN-2347-4890
Volume 4 Issue 1 January, 2016

Multipage Dynamic Apriori Approach for Web Log Mining

Author: Poulami Das[1] M.Tech. Scholar; Associate Prof. Nitin Mishra[2]
College: NRI Institute of Information Science and Technology[1], Bhopal; NRI Institute of Information Science and Technology1[2], Bhopal
Affiliation: Rajiv Gandhi Proudyogiki Vishwavidyalaya, India[1]; Rajiv Gandhi Proudyogiki Vishwavidyalaya, India[2]
E-mail:poulami.it@gmail.com[1]; nitin.nriist@gmail.com[2]

Abstract

Web Usage Mining is the process of extracting useful information from server logs. In order to produce server log patterns, this work will implements a high level process of Web Log Mining. Dynamic Improved Apriori Algorithm is one of the methods for the same that is used to finding out the association rules about a target page. But its efficiency is not high, as because, it finds out the association rules about a single target page only. Frequent patterns are patterns that occur frequently in a data set, which depends on minimum support threshold. This work finds out the frequent patterns in the area of Web Log Mining. This paper designs and implements a new algorithm called Multipage Dynamic Apriori Approach for Web Log Mining (MDAA-WLM), it works on more than one target pages and finding out all the frequent patterns about target pages based on minimum support count threshold. As compared to the Dynamic Improved Apriori Algorithm experiment shows that, the time of MDAA-WLM is improved in two areas- efficiency under different data volumes and efficiency under different minimum support count.

Keywords: web usage mining; web log mining; dynamic improved apriori algorithm; association rules; frequent patterns; multipage dynamic apriori approach;

1. INTRODUCTION:

Data Mining supports in doing an automated generating and extraction of anticipated information from the huge volume of data. The application of data mining technique is apply to the web is called web mining or web data mining. By the technique of web data mining, it can discover patterns and retrieves useful information from the World Wide Web services and documents. World Wide Web is an information system where one document is connected to the other document by hypertext links and user can search information by moving one document to the other document. Any requests of users in web is recorded in a file is called web log file. Extracting log files from the web is the process of web log mining. Web log mining can discover the relationship between user behavior and page content. Purposes of web log mining are: examine the recital of the web site; progress website design; to know user goal. There are three phases of web log mining: 1. preprocessing of log records, 2. discovering of patterns. 3. Analysis of patterns. [1, 2] Data preprocessing means to filtering, cleaning or eliminating outliers or irrelevant data from web log files. Integration of web usage data [3] from web server logs, corporate database, referral logs, registration files. There are different types of pattern discovery techniques. Out of them, mining frequent patterns, association rule and correlations are very important mining techniques. Mining frequent patterns leads to the discovering of interesting associations and correlations within data. Association rules mining is very common data mining task. This paper works on finding out all frequent patterns [4] from web log data. The analysis of pattern tool is like "which pages are being accessed most frequently?" the output of the analysis should be like "The frequency of visited per

IJournals

© 2016, IJournals All Rights Reserved

pages". There are three types of frequent patterns: frequent item sets, frequent subsequent, frequent substructure. This paper works on to finding out frequent item sets. Frequent item sets are those item sets whose support is equal or higher than the minimum support threshold. This threshold can be set by users or domain experts.

2. RELATED WORKS ON WEB LOG MINING:

E-Web Miner is a very effective tool in web log mining, which removes all flaws of apriori algorithm. In apriori algorithm, it takes time for repetitive scanning of database. By using E-Web Miner, the number of database scanning is hugely getting reduced. [5]

FP-growth algorithm is a very efficient and scalable algorithm for mining. In apriori algorithm, it takes more times, when it performs joining operation. By using FP-growth algorithm, it removes the joining step of apriori algorithm. For that reason, FP-growth algorithm is much faster than apriori algorithm. [6]

Based on web usage mining, it retrieves efficient web information by classifying users based on their internet usage patterns and for every class; it maintains a cache of web documents. [7]

In web log mining, it has three phases. [2] After completion of preprocessing pattern, the discovery of patterns and analysis of patterns plays an important role to identify the criminal activities and predict alleged user activities. [8]

Web log data is converted into sparse matrix and it calculate the influence degree of every web page for all web users to build a Matrix of Influence Degree. After that, cluster web users based on Matrix of Influence Degree. [9]

In internet there are some public information and some private information. Public information we can get easily. But to get private information, one Application Program Interface (API) is developed. In this way we can access private log files. [10]

Dynamic Data Mining is an approach, which can count item sets dynamically from a large amount of data. Dynamic Itemset Counting Engine (DICE) is a simple fast engine for counting item sets. It has three operations: 1.Add Item, 2.Delete Item and 3.Get Item sets. [11]

In real world, data mining is done dynamically. Database is changing over time. So, for that, there need to set different constraint to find out real informative rules. DMA-CO (Dynamic Mining of multi-supported Association rules with classification ontology) is an algorithm by which this work is performed very efficiently. [12]

The idea of argumentation reasoning is that, a statement is acceptable if it can be argued successfully against attacking arguments. By using PADUA (Protocol for Argumentation Dialogue Using Association Rules), it can dynamically mines Association Rules to generate the arguments exchanged among dialogue participants and represent every participants background domain knowledge. [13]

3. DYNAMIC IMPROVED APRIORI ALGORITHM:

This algorithm is a very influential algorithm which can mine pages dynamically. Apriori algorithm is static algorithm. But in actual data mining, data of database are always changing. This algorithm is used to finding out all the association rules for a single target page. But the efficiency of dynamic improved apriori algorithm is not high. Because, it is working only on a single target page and it takes more time too. For this reason a new algorithm is proposed called MDAA-WLM. This algorithm selects more than one target pages to finding out all frequent patterns for selected target pages. [14]

4. MULTIPAGE DYNAMIC APRIORI APPROACH:

This paper is introducing a new algorithm called MDAA-WLM, which is used to find out all frequent patterns for multiple pages. It will discover frequent patterns of all selected interested pages. In MDAA-WLM, it has extracted the knowledge of Negative selection algorithm. [15] Some principal and theorem of MDAA-WLM: Nature one: If one user is interested on one thing, then other user also may be interested on the similar thing. Nature two: If one user is interested on certain thing, then that user may be interested on similar things too. Nature three:

IJournals

© 2016, IJournals All Rights Reserved

Multiple users can be interested on multiple items at a time. First nature is based on collaborative filtering algorithm, second nature is based on content-based algorithm and third nature is the combination of both.

Theorem one: All nonempty subsets of a frequent item set must also be frequent. Theorem two: Superset of a non frequent item set cannot be frequent. Theorem three: Filter unique items from data set.

Web log is updated at any time. There is no need to generate all frequent patterns at one time, just real-time mining related frequent patterns for specific pages. In this paper, it finds out all the frequent patterns about more than one target page. Suppose you want to find out the frequent patterns about the pages 1, 2, 3, 4, 5. Suppose minimum support count threshold is 2.

1. Go through the web log database, statistics users who has visited target pages (for eg-1,2,3,4,5) as user set(U).Assume that U has 18 members-3,7,8,9,11,21,28,29,34,40,73,92,16,79,44 ,68,17,6.
2. Again go through the web log database and statistics visited pages for each user in U and get Page User Set(T), shown in (Table 1).It has total n records. n=37.

Table 1-Page User Set (T)

page	user	page	user	page	user
1	8	3	9	4	40
1	73	3	28	4	34
1	92	3	29	4	68
1	21	3	11	4	73
1	34	3	92	4	17
1	3	3	40	5	6
2	9	4	16	5	17
2	28	4	79	5	34
2	29	4	3	5	73
2	11	4	44	5	3
2	92	4	29	5	92
2	40	4	11		
2	34	4	92		

3. Create a new data arrangement in the form of page, user and num (number of user). Go across T statistical user set for each page set T, users are recorded to user attribute, and numbers of users are counted in num attribute. Here, we are not considering flag attribute. As because; these all pages are our target pages. So, flag is 1 for all pages, for this reason, in new algorithm, it needs not to consider flag attribute. Here minimum support count threshold is 2.We can get 1-element set (W1), shown in (Table 2). If the value of num attribute is less than 2, then this record should be deleted from W1.

Table 2-(W1)

Page	User	Num
1	8,73,92,21,34,3	6
2	9,28,29,11,92,40,34	7
3	7,28,29,11,92,40	6
4	16,79,3,44,29,11,92,40,34,68,73,17	12
5	6,17,34,73,3,92	6

4. Generate 2-element set (W2) from W1.This is joining step, shown in (Table 3). Whenever joining step is performed, that time it needs to be checked that, is there any common users, who are accessing both pages together.
5. Next is prune step. Delete record set, where num is less than 2, shown in (Table 3). Because our minimum support count threshold is 2. After deletion we can get next table (Table 4).

Table 3- Join and Prune Step (W2)

page	user	num
1,2	92,34	2
1,3	92	1(delete)
1,4	3,92,34,73	4
1,5	34,73,3,92	4
2,3	28,29,11,92,40	5
2,4	29,11,92,40,34	5
2,5	34,92	2
3,4	29,11,92,40	4
3,5	92	1(delete)
4,5	17,34,73,3,92	5

IJournals

© 2016, IJournals All Rights Reserved

Table 4- After Prune Step (W2)

Page	User	Num
4,5	17,34,73,3,92	5
2,5	34,92	2
1,5	34,73,3,92	4
3,4	29,11,92,40	4
2,4	29,11,92,40,34	5
1,4	3,92,34,73	4
2,3	28,29,11,92,40	5
1,2	92,34	2

6. Now generate 3 or more element set(W3) from W2, shown in (Table 5).
7. And so on, we can get 4 or more element set (w4), shown in (Table 6) and subsequent element set (if any).

Table 5-(W3)

Page	User	Num
2,4,5	34,92	2
1,4,5	34,73,3,92	4
1,2,4,5	34,92	2
1,2,5	34,92	2
2,3,4	29,11,92,40	4
1,2,4	29,11,92,40,34	2

Table 6-(W4)

Page	User	Num
1,2,4,5	34,92	2

This process will be continued up to there, where no more joining is possible. After completion of all steps, final result is shown in Table 7, where, all possible frequent item sets are showing dynamically about pages 1, 2, 3, 4, 5, which satisfies minimum support threshold.

Table 7- Final Result - All Frequent Patterns about Pages 1, 2, 3, 4, 5

Page	User	Num
4,5	17,34,73,3,92	5
2,5	34,92	2
1,5	34,73,3,92	4
3,4	29,11,92,40	4
2,4	29,11,92,40,34	5
1,4	3,92,34,73	4
2,3	28,29,11,92,40	5
1,2	92,34	2
2,4,5	34,92	2
1,4,5	34,73,3,92	4
1,2,4,5	34,92	2
1,2,5	34,92	2
2,3,4	29,11,92,40	4
1,2,4	29,11,92,40,34	2

5. EXPERIMENT AND RESULT ANALYSIS:

Datasets of 37248 records are used for our experiment. In the first experiment (Figure 1) it shows the difference in time according to different data volume 500, 1000 and 1500 for similar support count.

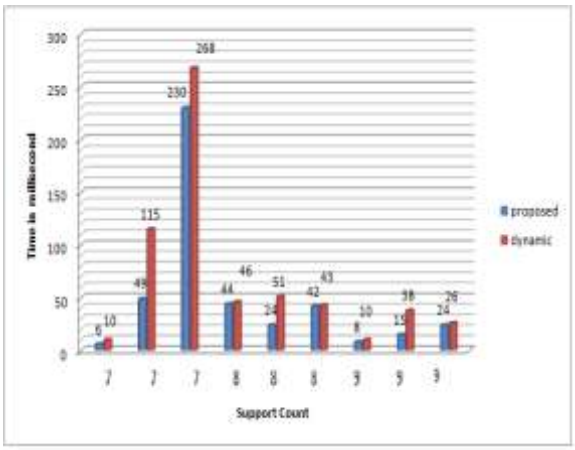

Figure 1: Comparison of algorithm efficiency under different data volumes

IJournals

In the second experiment (Figure 2) it shows that the difference in time of similar data set under different support count. Datasets are 3000, 4000, 5000 and 6000. For dataset 3000, support count is accordingly 10, 9, 8. Like this way others are also calculated. For both of the experiments it shows that, as compared to the simple dynamic improved apriori algorithm, MDAA-WLM shows very less time. So, time is improved in MDAA-WLM very effectively.

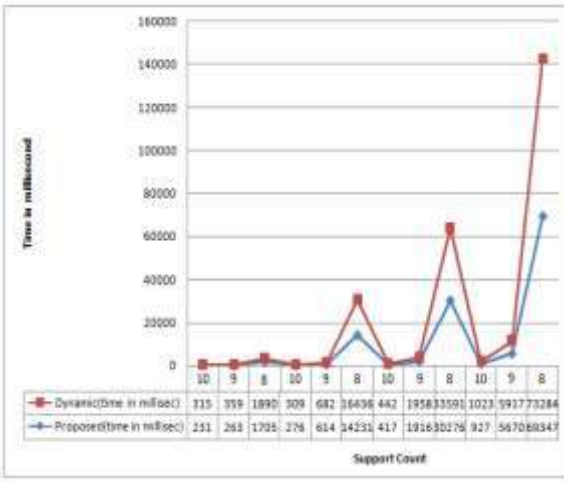

Figure 2: Comparison of algorithm efficiency under different support count

Hardware environment: Windows Edition- Windows 8.1, Processor- Intel Core I5, RAM- 2GB, System Type- 64 bit-Operating System . Software environment: Tools - Microsoft Visual Studio 2012, Platform - .NET 4.0, Language – C#.

6. CONCLUSION AND FUTUTE SCOPE:

This paper presents a frequent patterns discovery algorithm, which is suitable for web log mining. For multipage selection, it greatly reduces the search range. We can achieve real time updates and improve the efficiency of the algorithm. Selection of more than one target pages can avoid keeping of flag attribute as well as can avoid taking of unwanted pages. It has better efficiency in terms of time as compared to simple dynamic improved apriori algorithm. This work is useful for any organization which is having large amount of log data in server like University and Colleges. This work can further use in web sequence mining and clustered mining also. Based on our view, there are still several critical research problems that need to be solved. Still it is a very research demanding subject. It is very vigorous field for research & it will create new hopes in internet based commerce.

7. REFERENCES:

[1] Sisodia, D.S., Verma, S., "Web usage pattern analysis through web logs: A review", International Joint Conference on Computer Science and Software Engineering (JCSSE) ,Bangkok, pp-49-53, 2012 IEEE.

[2] Jianli Duan, Shuxia Liu, "Research on web log mining analysis", Instrumentation & Measurement, Sensor Network and Automation (IMSNA), International Symposium on (Volume:2), Sanya,pp-515-519, 2012 IEEE.

[3] K. Sudheer Reddy,G. Partha Saradhi,S. Sai Satyanarayana Reddy,"Understanding the scope of web usage mining & applications of web data usage patterns",pp-1-5,Dindigul, Tamilnadu, 22-24 Feb. 2012 IEEE.

[4] Dhaval S Patel, Jayraj M Desai, Swapnil Andhariya," A review on different frequent pattern mining techniques", International Conference of Advance Research in Computer Science and Management Studies, pp-23-26, Volume 3, Issue 2, February 2015.

[5] Mahendra Pratap,Pankaj Kumar Keserwani, Shefalika Ghosh, "An efficient web mining algorithm for web log analysis: E-Web Miner". Recent Advances in Information Technology (RAIT), 1st International Conference on Dhanbad, pp-607-613, 2012 IEEE.

[6] Mr. Rahul Mishra Ms. Abha 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,pp-311-318, Volume 2, Issue 9, September 2012.

[7] Indrajit Mukherjee, Samudra Banerjee,"Efficient web information retrieval based on usage mining". 1st International Conference on Recent Advances in Information Technology (RAIT), pp-591-595, 2012, IEEE.

[8] Amit Pratap Singh, Dr. R. C. Jain –" A survey on different phases of web usage mining for anomaly user behavior investigation" International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), pp-70-75, Volume 3, Issue 3, May-June 2014.

[9] Xiuming Yu, Meijing Li, Keun Ho Ryu-"Clustering of web users based on matrix of influence degree", 3rd International Conference on Applied Computing and Information Technology/2nd International Conference on Computational Science and Intelligence. pp-115-120, 2015 IEEE.

[10] Sachin Pardeshi, Tareek Pattewar, "Free user's behaviour information from central database system (web mining)", Intelligent Systems and Control (ISCO), 7th International Conference, pp-335-339, 2013 IEEE.

IJournals

[11]Sergey Brin and Lawrence Page,"Dynamic Data Mining: Exploring Large Rule Spaces by Sampling", pp 261-281, 1998.

[12] Ming-Cheng Tseng, Wen-Yang Lin, Rong Jeng," Dynamic mining of multi-supported association rules with classification ontology", Institute of Information Engineering Journal of Internet Technology, pp-1-8, march-11-2014 ResearchGate.

[13] M. Wardeh, T. Bench-Capon, F. Coenen,"Dynamic rule mining for argumentation based systems", Research and Development in Intelligent Systems XXIV, pp 65-78, 2008 Springer London.

[14] RuPeng Luan, SuFen Sun, JunFeng Zhang, Feng Yu, Qian Zhang,"A dynamic improved apriori algorithm and its experiments in web log mining". 9th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), Sichuan, pp-1261-1264, 2012 IEEE.

[15] Agnika Sahu, Prabhat Ranjan Maharana," Negative Selection method for virus detection in a cloud " Agnika Sahu et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, pp-771-774, Vol. 4 (6) , 2013.

IJournals

© 2016, IJournals All Rights Reserved