International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-8 Issue-6, August, 2019

Clustering Relatedbehaviour of Users by the Use of Partitioningandparallel Transaction Reduction Algorithm

C.Thavamani, A.Rengarajan

Abstract*: Fast improvement of data in relationship in the present universe of business trades, wide data getting ready is a primary issue of Information Technology. By and large, an Apriori figuring is extensively used to find the relentless thing sets from database. Later drawback of the Apriori count is overpowered by various estimations yet those are in like manner inefficient to find visit thing sets from far reaching database with less time and with amazing profitability. From this time forward another structure is proposed which contains facilitated passed on and parallel preparing thought. The examinations are directed to discover visit thing sets on proposed and existing calculations by applying diverse least help on various size of database. With expanded dataset, Apriori and Transaction decrease calculation gives horrible showing when contrasted with Partitioning and Parallel Transaction Reduction Algorithm(PPTRA). The actualized calculation demonstrates the better outcome as far as time intricacy and furthermore handle enormous database with more productivity.

Keywords* : Preprocessing, Mining of Association rules, frequent item sets, parallel, Apriori, matrix, minimum support, Partitioning.

I. INTRODUCTION

A gigantic proportion of research has been done on Web Usage Mining (WUM) which gives the information about the customer look for direct. At the point when the client peruses the pages, client leaves certain profitable data put away in the Web server get to log. This substance is extremely useful in deciding the web navigational example of client and the sort of data client needs from the particular sites.WUM includes essentially three noteworthy procedures to be specific information pre-treatment, design mining and example investigation by Udayasri.B, Sushmitha.N, Padmavathi.S. Right off the bat, Pre-treatment of information is done on an arrangement on Web logs to acquire logs with limited redundancies, client, session, exchange recognizable proof and data on way finishing. The pre-preparing errand is the initial phase in Web use mining, being in charge of perusing the web logs and prompting the relating client route sessions. All the while, the log information is cleaned so as to evacuate sections that are not helpful to demonstrate the client web route conduct and for fixing mistaken information. Client ID depends on data accessible in the log record, for example, the IP address, the kind of working framework and the perusing programming. Client route sessions are gotten from the log document. The sessionization assignment comprises of collection a succession of client's page demands into a unit named session. A session can be characterized as an arranged gathering of pages gotten tso by a client in a period window characterized by the minute the client entered the site and the minute the client left it. Also, mining calculations are connected to remove a client route design which speaks to relationship among Web pages in a specific Web website. In conclusion, design breaking down calculation is connected to separate information for information mining applications.. An Apriori calculation is broadly used to discover the incessant thing sets from database. An Association govern assumes a vital part in late information mining procedures. The obtaining of one item alongside another related item speaks to an affiliation run the show. Affiliation rules are utilized to demonstrate the connections between information things. Affiliation rules are as often as possible utilized for various purposes like promoting, publicizing and stock shop. Affiliation precludes discover normal utilization of things. This issue is persuaded by applications known as the market bushel investigation to discover connections between things by KeranaHanirex D, Dr.M.A.DoraiRangaswamy (2011). The relationship between information is entangled and the greater part of them is covered up. Affiliation principle mining is the for the most part utilized strategy in Association Knowledge Discovery which point is to discover the shrouded data. The most popular is the Apriori calculation which has been acquired 1993 by Agrawal, etl (1993). But it has two deadly bottlenecks:

(1) It needs great I/O load when frequently scans database.

(2) It may produce overfull candidates of frequent item sets.

To illuminate the bottleneck of the Apriori calculation by KeranaHanirex D, Dr.M.A.DoraiRangaswamy (2011), proposed framework will utilized PAFI (Partition Algorithm for Mining Frequent Item sets) for bunching and Matrix calculation to discover visit thing set from each group. This calculation segments the database exchanges into bunches. Bunches are shaped dependent on the similitude measures between the exchanges. Subsequent to framing the bunches we have to discover visit thing sets from each group utilizing lattice based technique by Feng WANG, Yong-huaLI (2008) with less measure of time. Henceforth the fundamental objective of the prescribed framework is to improve time intricacy and expectation precision.

Revised Manuscript Received on October 30, 2019.

* Correspondence Author

C.Thavamani*, Research Scholar, Bharathiar University, Coimbatore, Tamil Nadu, India.

A.Rengarajan, Professor, Department of CSE, Veltech Multitech Dr.RS Engineering College, Avadi, Chennai, Tamil Nadu, India.

II. LITERATURE SURVEY

Mining of frequent pattern is the mining of the frequently occurring ordered events or sub sequences as patterns. Now a day's many algorithms available to find the frequent item set from database.

Find frequent item sets using Apriori algorithm:

The most celebrated is the Apriori calculation which has been gotten 1993 by Agrawal which uses affiliation standard mining by Chen Wenwei (2006).

Affiliation principles are generally required to fulfill a client determined least help and a client indicated least certainty simultaneously. Affiliation rule age is normally part up into two separate advances:

1. Minimum help (limit) is connected to locate all regular thing sets in a database.
2. These continuous thing sets and the base certainty imperatives are utilized to shape rules.

Favorable position of this calculation, it is anything but difficult to discover visit thing sets if database is little however it has two savage bottlenecks. To begin with, It needs incredible I/O load when as often as possible outputs database and Second, It might deliver overfull up-and-comers of regular thing sets.

Find frequent item sets using PAFI as well as Apriori algorithm

D.KeranaHanirex, Dr..M.A.DoraiRangaswamy proposed productive calculation for mining incessant thing sets utilizing grouping systems. They displays an effective Partition Algorithm for the Mining Frequent Item sets (PAFI) utilizing grouping. This calculation finds the continuous thing sets by dividing the database exchanges into many bunches.It finds the incessant thing sets with the exchanges in the groups straightforwardly utilizing enhanced Apriori calculation which additionally diminishes the quantity of sweeps in the database and also simple to oversee and accessible effectively, consequently enhance the productivity and in addition new calculation superior to the Apriori in the space intricacy yet again it utilizes apriori calculation subsequently effectiveness not increment as much as required.

Find frequent item sets using Improved Apriori algorithm based on matrix

Feng WANG and Yong-hua proposed an improved Apriori calculation dependent on the framework. To explain the bottleneck of the Apriori calculation, they present an improved calculation dependent on the lattice by Zhu Yixia, Yao Liwen, Huang Shuiyuan, Huang Longjun (2006). It utilizes the network successfully show the undertakings in the database and utilizations the "AND task" to manage the lattice to create the biggest incessant thing sets and others. The calculation dependent on lattice don't filter database habitually, which decrease the spending of I/O. So the new calculation is superior to the Apriori in the time multifaceted nature however it isn't reasonable for huge database.Its understand that PAFI algorithm is better for partitioning large database and because of partition each cluster or partition easily swap in or swap out as well as Matrix method is better for find out frequent item set from each cluster with less span of time hence by using mixture of PAFI and Matrix based algorithm, it is easy to achieved frequent item set with better time and space complexity.

III. PROPOSED WORK

The principle purpose behind getting the proposed technique is, for an expansive Web website, there are just few individuals with comparable behaviorsand in this manner just these sort of regular examples are the most well-known ones among most of the clients by Suneetha.K.R, Dr.R.Krishnamoorthi (2009).

To explain the bottleneck of the Apriori calculation for example it needs incredible I/O load when much of the time examines database and it produces overfull applicants of continuous thing sets so it is trying to lessen the quantity of outputs their by diminishing the time and fundamental memory prerequisite.

Problem Definition

The most vital thought utilized is to diminish number of goes of exchange database outputs and psychologist number of applicants with the goal that it is effectively fit into principle memory regardless of the possibility that database is huge. Consequently to diminish the quantity of applicant it is proposed to, separate the entire database in to various group utilizing PAFI calculation After discovering the bunches, framework strategy for exchange lessening by Feng WANG, Yong-huaLI (2008) is connected on each group with the goal that it don't have to check database once more.

The architecture used in existing web usage mining framework is shown in**Figure1.** Some issues in the existing system are as follows

- No efficient data structures used for its data organization.
- No complete sessionization& preparation of data model based on Page views.
- Applied pattern mining techniques to the entire dataset includes both interested and not interested users which leads long execution time and use of more memory.
- Focused only on a specific/homogeneous data sets
- Use of Sequential search for pattern matching process leads to a worst case of time complexity.

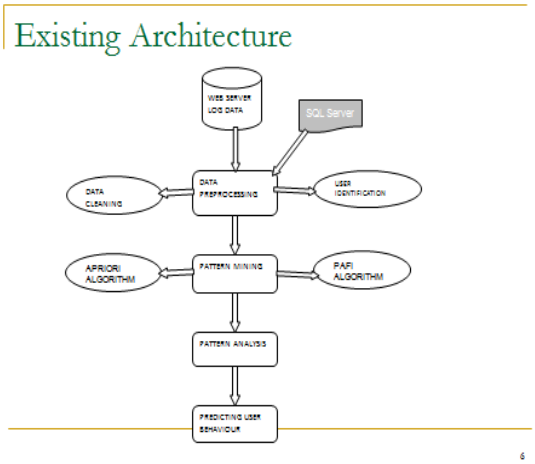

Figure.1 Existing Architecture Of Web Usage Mining

The New System Architecture

Proposed system uses two algorithms such as PAFI for clustering and then Matrix method on each cluster. In

proposed algorithm first, large database partition into different clusters to achieved better space complexity and then frequent item sets are found from all the clusters using matrix method for achieving better time complexity and thus it can overcome from both the drawback of apriori algorithm.In proposed algorithm combine two algorithms called PAFI and the Matrix based algorithm used. Below algorithm shows the steps of proposed algorithm.

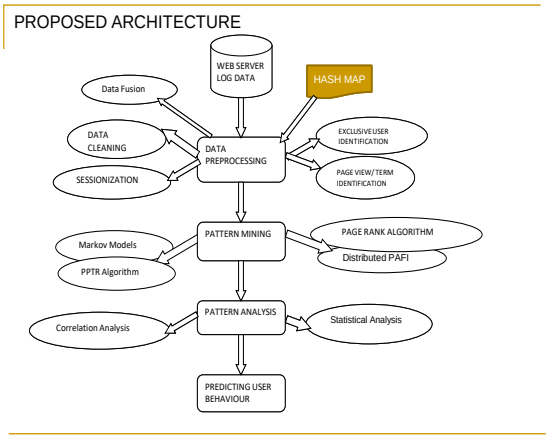

Figure.2 Proposed System Architecture Of Web Usage Mining

Algorithm:

Input: Database, Threshold and Number of clusters.
Output: Generate clusters, matrix and frequent item sets

Steps:-

1. Given set of transaction in the database.
2. Read Number of clusters.
3. Arrange all transaction in descending order, put it in the list.
4. As per input of number of cluster, select that many transactions in the list from the top and place it on the first position of every cluster.
5. After selection of first transaction in every clusters scan all transaction one by one and put highest similarity or minimum 3 similar items transaction in that cluster.
6. Step 5 will repeat till all transactions will be scanned.
7. Select next cluster from the list and repeat step 5 and 6.
8. Generate all clusters as per input.
9. Convert first cluster into matrix form.
10. After form the matrix, if the specific webpage on the transaction is available then mark 1 otherwise mark 0.
11. Find out all WebPages (K) of the matrix and find out the number of transactions (N) contains that web pages by AND operation.
12. If N>minimum threshold then K is a frequent item set.
13. Then consider different combination of K-1 item a much as possible using Markov models.
14. Go to step 13 till find all frequent item sets in that cluster.

Now take next matrix in to hash map memory area and repeat steps 10 to 15 till get frequent item sets from all matrices.

In this algorithm, number of clusters, number of minimum similar items from transaction and minimum support threshold is decided by user depends on the applications. After that the entire database divided into that many clusters. After generating the cluster, the clusters that have the total number of transactions less than some threshold value will be deleted.

Now it is easy to apply matrix algorithm on each cluster rather than applying matrix algorithm on entire database. Cluster will required less space hence memory complexity also increases. It is easy to find out frequent item sets from cluster than entire database.After applying matrix algorithm on each matrix, generate FIS (frequent item set from all matrix (all clusters) and arrange frequent item set of all cluster in to the array.

IV. EXPERIMENT RESULTS

Ideally, the input for the Web Usage Mining process is a web server log file which contained raw data of user interactions with many web pages as shown in Table

Table.1 NASA Web Server Log File

S.No	Log details
1	199.72.81.55 - - [01/Jul/1995:00:00:01 -0400] "GET /history/apollo/ HTTP/1.0" 200 6245
2	199.120.110.21 - - [01/Jul/1995:00:00:09 -0400] "GET /shuttle/missions/sts-73/mission-sts-73.html HTTP/1.0" 200 4085
3	199.120.110.21 - - [01/Jul/1995:00:00:11 -0400] "GET /shuttle/missions/sts-73/sts-73-patch-small.gif HTTP/1.0" 200 4179
4	205.212.115.106 - - [01/Jul/1995:00:00:12 -0400] "GET /shuttle/countdown/countdown.html HTTP/1.0" 200 3985
5	129.94.144.152 - - [01/Jul/1995:00:00:13 -0400] "GET / HTTP/1.0" 200 7074
6	129.94.144.152 - - [01/Jul/1995:00:00:17 -0400] "GET /images/ksclogo-medium.gif HTTP/1.0" 304 0
7	199.120.110.21 - - [01/Jul/1995:00:00:17 -0400] "GET /images/launch-logo.gif HTTP/1.0" 200 1713
8	205.189.154.54 - - [01/Jul/1995:00:00:24 -0400] "GET /shuttle/countdown/ HTTP/1.0" 200 3985
9	205.189.154.54 - - [01/Jul/1995:00:00:29 -0400] "GET /shuttle/countdown/count.gif HTTP/1.0" 200 40310
10	205.189.154.54 - - [01/Jul/1995:00:00:40 -0400] "GET /images/NASA-logosmall.gif HTTP/1.0" 200 786
11	205.189.154.54 - - [01/Jul/1995:00:00:41 -0400] "GET /images/KSC-logosmall.gif HTTP/1.0" 200 1204

A page view is an aggregate representation of a collection of Web objects to the display single user action (such as a click-through). Each page view can be viewed as a collection of Web pages or resources representing a specific "user event". Eg., Reading an article, viewing a solar system page, or apply for NASA jobs.

Data pre-processing produces

A set of page views: P={p_1, …, p_n}

A set of user transactions: T={t_1, …, t_m}

where each transaction $t_i$contains a subset of P. Each transaction t includes page view and its weight.

$$t = \{(p_1^t, w(p_1^t)),(p_2^t, w(p_2^t)),\ldots,(p_l^t, w(p_l^t))\}$$

is a l length ordered sequence of page views, where each w corresponds to a weight. Weight represents the duration of the page visit in the session.

$$t = (w_1^t, w_2^t, \ldots, w_n^t)$$

where the weight is 0 if the corresponding page is not present in t, otherwise correspond to the significance (weight) of the page in the t. For a given set of transactions in the database D,which consist of only 9 transaction and 5 items and it divided into two clusters. After forming cluster using PAFI algorithm, now apply the transaction reduction algorithm (matrix) on each cluster i.e. CL1 and CL2 but here CL2 has less number of transactions that is less than the threshold value so we are deleting the transactions in CL2 and applying transaction reduction algorithm only on the transactions in CL1.

This representation of CL1 is shown in**Table.2**.

Table.2 Transaction Matrix/ User-Page view Matrix (M×N)

Users/Page Views	Page View A	Page View B	Page View C	Page View D
User1/T1	15	4	0	1
User2/T2	2	0	25	0
User3/T3	20	1	0	0
User4/T4	56	0	0	1
User5/T5	0	0	23	55

Find out the Matrix of CL1

The term-page view matrix represents the concepts that appear in each page like space, earth, mars, aeronautics, solar etc. The weights in the term-page view matrix are usually a function of specific term frequencies usually represented as binary values shown in **Table.3**.

Find out the largest frequent itemsets

We aim to transform a user-page view matrix (MxN) and Page view-Term view matrix (NxR) into a content-enhanced transaction matrix by doing AND of (MxN) , (NxR) gives (MxR) shown in **Table.4**.

Table.3Term- Page View Matrix (N×R)

Page views/Terms	Space	Earth	Mars	Aeronautics	Solar
Page View A	0	1	0	1	0
Page View B	1	1	1	0	1
Page View C	1	0	0	0	1
Page View D	0	1	0	1	0

Table . 4Content Enhanced Transaction Matrix (M×R)

Users/Terms	Space	Earth	Mars	Aeronautics	Solar
User1	1	3	2	2	2
User2	1	1	0	1	1
User3	1	2	1	1	1
User4	0	2	1	2	1
User5	1	1	0	1	1

By clustering the rows of the above matrix may reveal users with common interests[12]. In the resulting matrix users 2 and 5 are more interested in concepts related to space and earth while user 1 is more interested in earth and mars. This helped us to identify users with common interests which can be used for further pattern matching, pattern discovery and analysis of common patterns.

Statistical Study

The experiment is conducted on dataset, which composed of more than 1000 transactions and average size of transaction is 5 web pages and based on that performance is measured with different parameters.The performance measured on different set of transaction with fixed threshold =3 is shown in Table 12 and figure 3.It shows that matrix and apriori algorithm is required more time when transaction size is increased as compared to PAFI with apriori and PPTRA.

Table.5 Time required to generate frequent item set with threshold = 3 on different algorithm.

No. of Transaction	Apriori (in Sec)	Matrix (in Sec)	PAFI with Apriori (in Sec)	PPTRA (in Sec)
100	**10**	**23**	**7**	**5**

200	456	302	16	9
300	992	1036	28	15
400	3189	4120	97	41
500	10349	12657	239	74
600	23890	26534	1253	158
700	58672	57249	2802	221
800	70213	68126	6544	307
900	87890	89343	12472	416
1000	93245	97128	25513	562

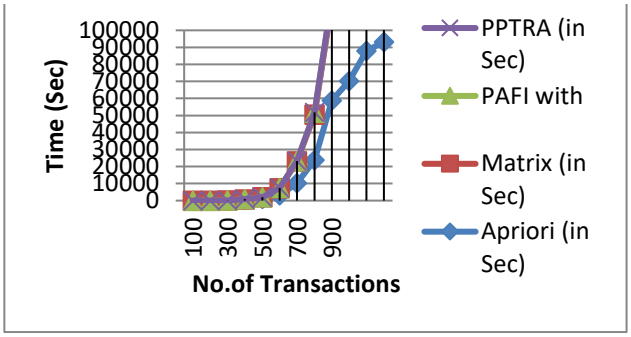

Figure.3: Time required by different algorithms on different set of transactions.

The performance measured on fixe dataset and with different threshold is shown in Table.6 andFigure4. It also shows that matrix and apriori algorithm is required more time when threshold is decreased as compared to PAFI with apriori and PPTRA.

Table .6 : Time required finding out frequent item set with 1000 transactions.

Threshold	Apriori (in sec)	Matrix (in sec)	PAFI with Apriori (in sec)	PPTRA (in sec)
3	93245	97128	25513	562
4	64345	76345	2702	559
5	45246	54389	961	485
6	20367	39455	705	482

PPTRA as well as PAFI with apriori both algorithm uses clustering technique hence the performances is measured on both algorithms with different thresholds and with different size of data set. It shows that PAFI with apriori gives faster FIS when number of transaction less than 500 and threshold =5 but when the transaction increases it becomes slower than PPTRA algorithm.

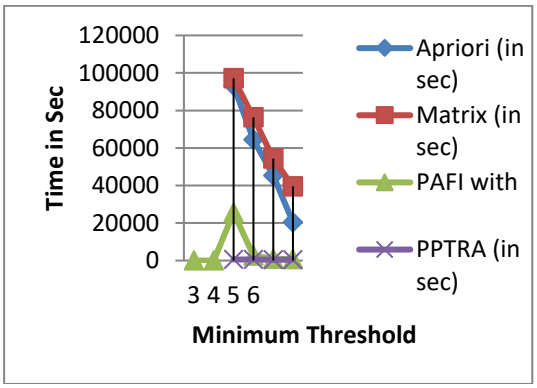

Figure. 4: Time required by different algorithms with different threshold.

When Number of transactions is less than 500 and threshold is 6 than PAFI with apriori work faster than PPTRA but as the threshold value decreases and number of transaction increases PPTRA is faster than PAFI with apriori.

V. CONCLUSION

In this paper, the novel computation PPTRA is proposed where the entire database isolated into assignments of variable sizes, each section will be known as a bundle than each gathering is changed over into structure by matrix count and make visit thing set from each gathering. Here Instead of entire database simply each gathering is seen as everyone thusly hereafter time required to swap in and swap out from memory is less appear differently in relation to apriori and Matrix figuring and furthermore computational speed will be addition. It similarly reduces the monotonous database look at and upgrades the productivity.Performance studies demonstrates that PPTRA take half to 80% less time than PAFI with apriori calculation to create FIS just as if limit worth changes on same dataset than additionally PPTRA take practically same measure of time while existing calculation shifts regarding change in edge esteem. It additionally demonstrates that Matrix and apriori isn't proficient for huge dataset.Hence tale calculation PPTRA gives preferred execution over existing calculations when there is huge dataset and it gives better time intricacy and space complexity.With the assistance of recognized examples and report of investigation we can improve the forecast precision of clients conduct.

REFERENCES

1. Agrawal R, Imielinski T, Swami A (1993), "Mining association rules between sets of items in large databases". In: Proc. of the l993ACM on Management of Data, Washington, D.C, May, 207-216.
2. Arvind K Sharma and P.C. Gupta (2013) , "Predicting the Behavior and Interest of the Website Users through Web Log Analysis", International Journal of Computer Applications, Vol. 64, No. 7.
3. Chen Wenwei (2006), "Data warehouse and data mining tutorial". Beijing: Tsinghua University Press.
4. Feng WANG, Yong-huaLI (2008):"Improvedapriori based on matrix",IEEE , 152-155.

5. Han Jiawei, KamberMiceline. Fan Ming, MengXiaofeng translation (2001), "Data mining concepts and technologies". Beijing: Machinery Industry Press.
6. KeranaHanirex D, Dr.M.A.DoraiRangaswamy (2011):" Efficient algorithm for mining frequent item sets using clustering techniques." In International Journal on Computer Science and Engineering Vol. 3 No. 3 , 1028-1032 .
7. Margatet H. Dunham(2003). Data Mining, Introductory and Advanced Topics: Upper Saddle River, New Jersey: Pearson Education Inc.
8. Suneetha.K.R, Dr.R.Krishnamoorthi (2009), "Data Preprocessing and Easy Access Retrieval of Data through Data Warehouse", WCECS, October 2009, 20-22, San Francisco, USA.
9. TongQiang, Zhou Yuanchun, Wu Kaichao, Yan Baoping (2003), "A quantitative association rules mining algorithm". Computer engineering. 33(10):34-35.
10. Udayasri.B, Sushmitha.N, Padmavathi.S , "A LimeLight on the Emerging Trends of WebMining" ,Special Issue of International Journal of Computer Science & Informatics (IJCSI), ISSN(PRINT): 2231–5292,Vol-II,Issue-1,2 .
11. Wael A. AlZoubi, Azuraliza Abu Bakar,Omar (2009)," Scalable and Efficient Method for Mining Association Rules", International Conference on Electrical Engineering and Informatics.
12. Zhu Yixia, Yao Liwen, Huang Shuiyuan, Huang Longjun (2006), " A association rules mining algorithm based on matrix and trees". Computer science, 33(7):196-198.
13. C. J. Kaufman, Rocky Mountain Research Lab., Boulder, CO, private communication, May 1995.
14. Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interfaces(Translation Journals style)," IEEE Transl. J. Magn.Jpn., vol. 2, Aug. 1987, pp. 740–741 [Dig. 9th Annu. Conf. Magnetics Japan, 1982, p. 301].
15. M. Young, The Techincal Writers Handbook. Mill Valley, CA: University Science, 1989.
16. (Basic Book/Monograph Online Sources) J. K. Author. (year, month, day). Title (edition) [Type of medium]. Volume(issue). Available: http://www.(URL)
17. J. Jones. (1991, May 10). Networks (2nd ed.) [Online]. Available: http://www.atm.com
18. (Journal Online Sources style) K. Author. (year, month). Title. Journal [Type of medium]. Volume(issue), paging if given. Available: http://www.(URL)