E-COMMERCE WEB MINER WEB MINING ALGORITHM FOR WEB LOG ANALYSIS

[1]Miss Rohini Sidhaling Patil , [2]Prof. Y.R.Kalshetty
[1]Student, [2]Assistant Professor
[1]Computer Science
[1]Shri Vithal Education & Research Institute's College Of Engineering, Pandharpur, India

Abstract—Now a day's electronic gadgets are easily available due to their affordable rates and drastic increase in E- commerce sites online shopping business industry reaches sky limits. So, millions of transactions are happening at the E-commerce site web servers. And the data dumps at the server logs will be in terms of gigabytes or sometimes it is in Terabytes. So, this much huge data is always plays an important role in identifying missing transactions or mining most reliable sold items which can be helpful for the E commerce industry. Many algorithms and methodologies are existed to mine the web log data like Apriori and others. But most of the algorithms are suffered from space complexity issue with the limited items sets. So, proposed system introduces a technique of mining the E-commerce web server logs using M- tree based frequent pattern analysis. And this process is powered with the Shannon information gain technique to identify the most important items that are distributed over the datasets.

Index Terms— Shannon information gain, M- tree, Web logs, Pre-order traversing, Support clustering.

I. INTRODUCTION

Web Server Logs are log file that maintains all the records or history of all the hits or requests that the server receives. When a user sends the request from the client computer to the server computer for individual files, the web server records all these requests in the form of logs. All the information received from the client computer is stored secretly and have all the detail information such as at what time and date the request is made from client side, the IP address of the request, the URL of the request etc. These web server logs are helpful during technical auditing and problem solving of the web server. These web server logs are not available to common internet users only the administrator has a right to access these logs.

Examination and analysis of web server logs help administrator to efficiently handle web site administration and resource hosting. Any organizations that are in the field of sale and have website, their marketing department should be trained to understand the web server logs, to make the sales effort smooth. Analysis of web server logs manually or with software package help in understanding user interaction with a website.

Frequent itemset Mining is most popular and important field of data mining, which initially applied in market basket analysis. These types of sequential analysis of actions or event sequences are very helpful in different areas to examine behavioral patterns e.g. in games. In frequent itemset mining the data get the set of instances or transactions which contain lots of features or items. After getting all the sets, the frequent itemset mining algorithm is used to categorize all the common sets of items, that mean the item sets having minimum supports or least minimum times it exists.

After data of frequent item sets identified, association rule can be implemented to analysis of behavioral patterns. It helps in finding customers shopping behavior in supermarket, online shopping etc. It helps in identifying the products the customer regularly purchased.

Once identified the sets of associated products the organization makes the best use of identifying products to make these products available on the supermarket shelves or online web pages. Frequent itemset mining is also used in games to identify the playing technique that frequently co-occurs between players. It is used in different fields and a variety of tasks for finding similarity between variables of the given data sets.

Frequent itemset mining is the most efficient data mining technique to identify the correlation and the relation among larger data transactions. E-commerce applications generate huge amount of transactions and operational data in e-commerce applications all the information about a customer from entry to exist of the shop is recorded.

Web server logs are an important way to collect e-commerce data, because most of the data mining technique depends on web server logs. Applying frequent itemset mining and association rule to these huge e-commerce application data to examine behavioral pattern and finding important information from these data.

So, customer finding patterns and purchasing behavior are identified by examining these e-commerce data. Finding association rules of these e-commerce data is helpful to make important decision in favor of business. It's also helpful, in selling products and providing purchasing suggestions to customers.

Frequent item set mining and association rules implementation in e-commerce data is used to personalize and recommend customizing solution to the customer, thus satisfying their personal needs by seeing the feedback from other user that already purchase the products. Frequent item set mining of e-commerce data also suggest useful and interesting products to the customers. Thus, overall frequent itemset mining is used to improve e-commerce application.

In a data structure, the tree is describing as a collection of nodes, starting with the parent node (root) together with the list of child nodes, and each node has a value assigned to them. There are different types of tree data structures, Binary Tree structure is one of them. In binary tree structure, every parent (root) has at last two children, referred as left child ("left sub-tree") and right child ("right-sub tree"). Each node in a binary tree is connected with other node by a directed edge. Each parent node is connected to a random number of nodes, called as child node. Nodes without child node are defined as leaves and nodes without leaves are defined as internal nodes.

A binary tree is called the "Perfect Binary Tree", if all interior nodes consist of two child nodes and all leaves are on the same level. A binary tree, where every node is having two or zero children, is called as "Full Binary Tree". In a "Complete Binary Tree", every level is completely filled and in the last level, each node is to the left side as possible. In a computer science binary tree is used in many applications where data are constantly entering and leaving.

In this paper, section 2 is dedicated for literature review of past works. Section 3 describes the proposed methodology and Section 4 discusses the results and evaluation of the proposed technique. Finally, Section 5 concludes this paper with future extension possibilities.

II. RELATED WORK

Scholar **[Yang]** work focuses on caching web objects to enhance performance of web caching systems. N-gram based algorithmic procedure is been developed to predict future request .GDFS caching system is at core with association rule mining. Limitation observed is system cannot handle dynamically changing database. Scope of work is design algorithmic procedure for query based matching.

Author **[Ivancsy]** pattern mining is been highly researched are in data mining with numerous applications. One major Application of frequent items is pattern identification using web log discovery. Web log mining assist in web discovery to identify user search behavior. This research would be extended in web user profiling. Three patterns mining Approaches have been studied page set, page graph and page sequence. Hidden data patterns have been identified using rule based data mining. Major limitation observed is overall algorithmic procedure is complex and requires higher hardware resources also.

[Pi-lian] Conventional search engines face challenge to apply data mining for log analysis. Future candidate set generation procedure is been applied with apriori algorithm. Proposed algorithmic procedure is lesser complex as compared to existing procedure. Overall memory requirement of Algorithm is also low. Future scope of work is categorizing only relevant logs and effective data processing.

Research presentation **[Borgelt]** available on Google scholar introduces readers with concept of frequent item set mining. Item set mining is Market Analysis process, aimed to profile user shopping behavior. Simple find products that are been buyed together. Association rule have commonly used in pattern mining. Finding common and uncommon purchasing patterns. In frequent item set transaction monitoring is been done to collect user data. Numerous approaches exist in generating common frequent item sets patterns. Commonly adopted is brute force traversal. Common challenge observed is effective pattern generation for fining frequent items. Existing Algorithm approaches like apriori fail to address common issues as such scope of work is development of better algorithmic procedure for frequent item set mining.

Article **[Grahne]** focuses to solve faster rule mining in generation of frequent item set. Proposed approach is FP growth algorithm implemented using Prefix tree data structure. Challenge is to traverse tree effectively. FP-Array mechanism is been presented for effective tree traversal. Even though algorithm is comparatively faster as compared to existing procedures .future scope of work remains in memory usage reduction. . Furthermore FP-tree is too large to fit in memory, the current solutions need a very large number of disk I/O for reading and writing FP-trees onto secondary memory or generating many intermediate databases making mining frequent item sets too time-consuming. Time reduction and large database scaling are scope of work.

Author **[Dong]** Address frequent item set generation in cloud, where privacy remains major challenge. Server results need to be verified as is challenge task for weaker client. In order to overcome this probabilistic and deterministic cloud data verification approach is been proposed. This framework assists in verification that returned frequent items set are correct. Existing system used interactive proof methodology for handling which is time and cost consuming procedure. Applying cryptographic mechanism became too odd due to this. as result validation remains unaddressed ,proposed research is only of one to address correctness of frequent item set. Future scope of work is to allow client in making specific verification and not complete verification.

Introduction and overall concept of information gain is been presented in [Ravi] in detailed with mathematical equations. Shannon theory has been developed on h Boltzmann' mathematical principle discrete probability space and finite elements in space are been described. gain theory finds the overall gain of knowledge obtained from set of elements in space.

Focuses to mine **[Koh]** rare patterns from medical, shop data and numerous data mining applications. Algorithmic strategies have been classified in aprio and tree based. This article highlights on rare patters in data mining. Association based rule mining is most common technique used in detecting rare patterns. Open challenges highlighted are pattern mining in streams, discarding noisy data, correct frame selection in data mining, real time data mining, mining patterns on rare dataset.

Research work of **[Yuan]** focuses on Among mining calculations in light of affiliation tenets, Apriori technique, mining incessant itermsets and intriguing relationship in exchange database, is not just he initially utilized affiliation administer mining strategy additionally the most famous one. In the wake of contemplating, it is discovered that the traditional Apriori calculations have two noteworthy bottlenecks: filtering the database every now and again; producing an expansive number of applicant sets.

In view of the characteristic deformities of Apriori calculation, some related improvements are completed: utilizing new database mapping approach to abstain from filtering the database more than once; additionally pruning successive itemsets and hopeful itemsets with a specific end goal to enhance joining proficiency utilizing over procedure to tally support to accomplish high productivity

Under similar conditions, the outcomes delineate that the proposed enhanced Apriori calculation enhances the working efficiency contrasted and other enhanced calculation algorithms. In scope for future work this made strides calculation can be additionally made strides from the viewpoint of space intricacy.

In application, it can be connected in medicinal services procedures to examine applicable restorative procedures and distinguish the compelling approaches to enhance the effectiveness of medicinal systems and administrations, keeping in mind the end goal to spare restorative costs and reduce the healing center weight for patients stuffing, and enhance patients' fulfillment.

Focus of **[Pinjia]** is on effective log processing algorithmic procedure. Log file record dynamic runtime information of user in form of records and are saved on servers. These files have been frequently mined by developer and programmer for effective process. Size of log files are ever increasing from mb to gb and as such effective data processing technique required. log parsing remains major task. Design and development of better log parser is open task in log mining.

Six log parsers have been studied and examined in detailed here. Real work data effectiveness of parser has been demonstrated. Future scope of work is designing source code parser.

[Jing] Research presents cluster formation in web logs. he benefits of FCM calculation are that it is for the most part connected in point information group and can't specifically handle social information, for which the paper proposes a grouping algorithm in information mining in light of web log. Right off the bat, the paper improves. experiments demonstrate that mining aftereffects of CA-FCM calculation is near the mining aftereffects of FCM calculation, and the execution of CA-FCM calculation is superior to that of FCM calculation when the number of client's access to session is not expansive. Scope of work remains effective evaluation of algorithm based on correct evaluation parameters.

[Dharmarajan] focuses on web mining to examine user behavior on online portals. This process would enhance web recommendation and better system feasibility Fp growth algorithmic strategy has been applied in data mining. This procedure provides vital information about user using hidden pattern discovery process.

The research could be extended to adaptive we recommendation, browser effective suggestions. In future system could be extended in image content mining from logs which is yet not been done. Better technique needs to be designed for effective repetitive research.

In **[Wang]** Network data log mining is proposed using fuzzy clustering process. Three-layer algorithmic procedure has been introduced preprocessing, pattern mining, pattern examination. Attacker time, attack process and network protocols have been examined from logs. Collaborative optimization has been done to enhance process of data log mining. Algorithm has found achieve better clustering results and higher efficiency.

Huge data has been generated in log **[Fernandez].** Effective data processing mechanism has been required for data analysis. Web logs need to be examined for effective expert analysis. Preprocessing methodology has been found to be best process in eliminating unwanted data from huge information. Effective patterns have been generated from data for search analysis.

Association rule based on fuzzy rule enhances better decision making. Personalized ontology model enhances system performance. As such ontology with preprocessing and data cleaning task is best approach in handling huge data. Scope of the work remains in design of personalized ontology model, dynamic updating of ontology associations.

III. PROPOSED METHODOLOGY

Overview of proposed methodology of web log mining for frequent item sets can be viewed on figure 1.

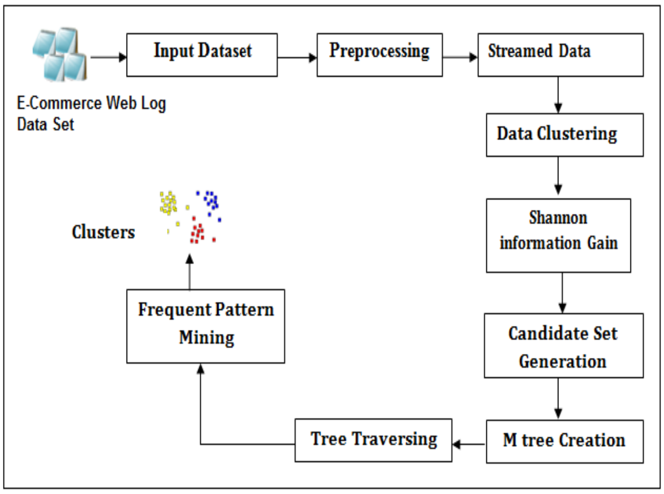


Fig 1: Overview of our approach

The steps of the proposed methodology of web log mining to identify the frequent item sets can be broadly analyze and understand using the following steps.

Step 1: Preprocessing - This is the initial step of the system where system is feeding with the web logs of E –Commerce data. This web log data set is collected from URL http://recsys.yoochoose.net/challenge.html . Here this data set consists of some attributes like Session ID, Item ID, Product ID and date time.

In this phase of the proposed model data set is read in the form of string and CSV format of the data is being converted into formatted vector of double dimension list for further processing of the system.

Step 2: Here in this segment of the proposed methodology important attributes has been identified using Shannon information gain technique.

To begin with this process proposed methodology first identifies the unique attributes from the Session ID, Item ID and price attributes using hashing of sets. After this process, all the attributes are separately clustered linearly with block size of 10 to evaluate the distribution gain of the information.

Then every unique entity of all the individual attributes are measured for their appeared frequency in the linear cluster to get the information gain value in between 0 and 1. Elements whose information value is nearer to 1 is consider as most distributed and so it is important.

This is achieved by using Shannon information gain theory which can be represented with the help of the following equation 1.

$$IG(E) = - (P / T) \log (P / T) - (N / T) \log (N / T) \quad \text{--------(1)}$$

Where

P= Frequency of the present count
N= Non-presence count
T= Cluster Elements Size.
IG(E) = Information Gain for the given Entity

Based on these values of the Shannon information gain top 4 Session Id’s , Top 3 items and top 3 Price entities are selected to create frequent item sets.

Step 3: Frequent item set Creation – Frequent item set has been created using apriori algorithm as Shown in algorithm 1 for the top items extracted from the prior step.

Algorithm 1: Apriori Algorithm for Frequent Item Generation

```
Input: Termset $K_d$ and support threshold S0
Step 1.Scaning Items F={Ø}
Step 2:for(i=1; i< 2^ I; i++)
$T_{count}$ =0;
Step 3:for(j=1; j<2^I; j++)
            If Ai $\leq$ Aj
            $T_{count}$t = $T_{count}$ + $S_{count}$ (j);
Step 4:If($T_{count} \geq$ S0)
          Then F = F $U$ Ai ( Candidate Set Generation)
Step 5: Goto step 2
Step 6: End
Output: Frequent itemsets
```

Step 4: M –Tree Creation – once the frequent item sets are generated along with their supports then they are processed in a double dimension list. This double dimension list is used to create M –tree. Here in this process first encountered candidate set is considered as the root. And then subsequent candidate sets are assigned their position based on their support value.

If the support values of the candidate sets are lesser than that of root node then they are assigned to the left of the root node or else right of the root node. This process continuously run in recursive manner to generate a well-mannered sorted tree called as M- tree and is depicted in Algorithm 2.

Algorithm 2: M- tree

```
//input : Candidate Set List $C_L$
Step 0: Start
Step 1: Create an empty tree as **T**
Step 2: Create the Root Node for first frequent itemset $\mathbf{R_n}$
Step 3 :**FOR** i=0 to size of $\mathbf{C_L}$
Step 4: Compare the distance with the root node $\mathbf{R_n}$
Step 5:If ($\mathbf{C_{Li}}$support <$\mathbf{R_n}$)
Step 6:Add node as left child in **T**
Step 7: Else
Step 8:Add node as Right child in **T**
Step 9:End **FOR**
Step 10: return **T**
Step 11: Stop
```

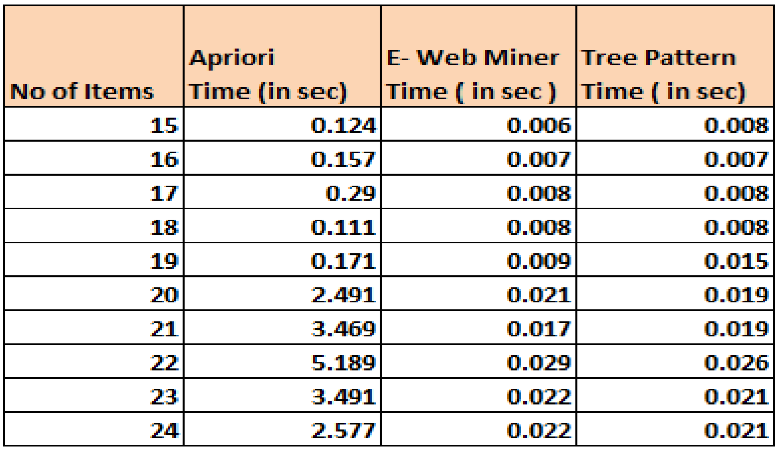
Step 5: This is the last phase of our model where all the candidate sets are traversed in pre-Order manner. Where nodes are traversing in ROOT, LEFT CHILD and then Finally RIGHT CHILD manner to collect the similar support candidate sets to form the clusters of the frequent item sets based on the similar support.

IV. RESULTS AND DISCUSISONS

To deploy the proposed idea in java technology system uses NetBeans 6.9.1 as the IDE. System uses the web server log data set collected from URL http://recsys.yoochoose.net/challenge.html.The collected dataset is in the .txt extension.

On its complete turn of the system, proposed model yields the clustered frequent item sets for the E- commerce data which reveals the best possible correlation between the entities.

To measure the performance of the system proposed methodologies results are compared with that of E-web miner and Improved Apriori Algorithm as stated in [16]. The results of the Experiments are tabulated as shown in table 1.



No of Items	Apriori Time (in sec)	E- Web Miner Time (in sec)	Tree Pattern Time (in sec)
15	0.124	0.006	0.008
16	0.157	0.007	0.007
17	0.29	0.008	0.008
18	0.111	0.008	0.008
19	0.171	0.009	0.015
20	2.491	0.021	0.019
21	3.469	0.017	0.019
22	5.189	0.029	0.026
23	3.491	0.022	0.021
24	2.577	0.022	0.021

Table 1: Experimental Results

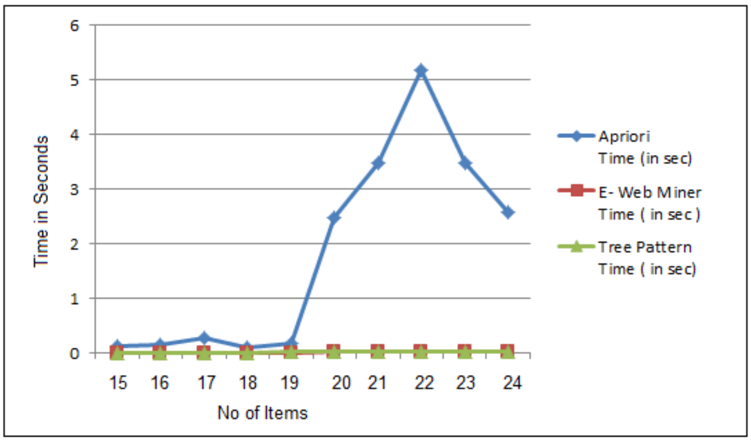


Figure 2: Comparison of all three methodologies

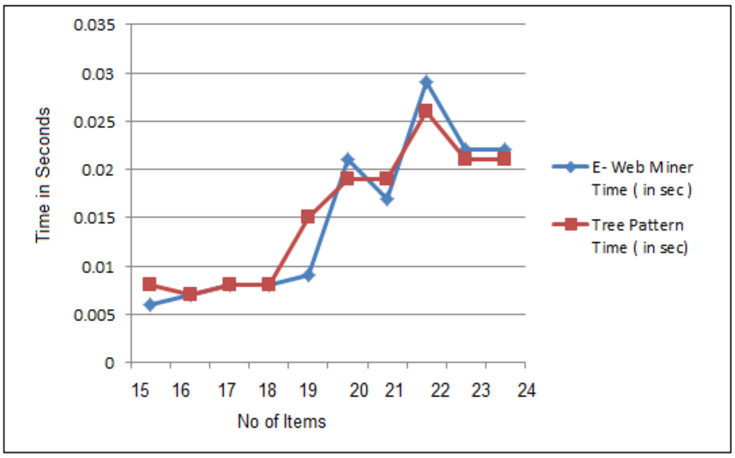


Figure 3: Comparison with E- web miner

On observing figure 2 we come to know that Apriori algorithm is struggling with is its performance time factor. Whereas on observing figure 3 we come to know that our methodology of frequent itemset mining using tree pattern technique over performs with thin line edge than that of web miner algorithm.11.

V. CONCLUSION AND FUTURESCOPE

Majority of the frequent itemset mining algorithms are yielding desired candidate sets based on the support. But most of them are not categorized them based on the support on their evaluation process.

Proposed model clusters the frequent item sets based on their support on the process of its formation using M-tree technique for the important and most distributed candidate sets decided by Shannon information gain. On evaluation of the system's performance based on time factor proposed methodology outperforms than that of E – web miner and Apriori algorithms as elaborated in the prior step.

In the future, this system can enhance to work on huge dataset with the distributed paradigm.

REFERENCES

[1] Yang, Qiang, and Henry Haining Zhang. "Web-log mining for predictive web caching." *IEEE Transactions on Knowledge and Data Engineering* 15.4 (2003): 1050-1053.

[2] Iváncsy, Renáta, and István Vajk. "Frequent pattern mining in web log data." *Acta Polytechnica Hungarica* 3.1 (2006): 77-90.

[3] Pi-lian, WANG Tong1 HE. "Web log mining by an improved Aprioriall algorithm." *Engineering and Technology, þ* 4.2005 (2005): 97-100.

[4] Borgelt, Christian. "Frequent item set mining." Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 2.6 (2012): 437-456.

[5] Grahne, Gösta, and Jianfei Zhu. "Fast algorithms for frequent itemset mining using fp-trees." IEEE transactions on knowledge and data engineering 17.10 (2005): 1347-1362.

[6] Dong, Boxiang, Ruilin Liu, and Hui Wendy Wang. "Trust-but-Verify: Verifying Result Correctness of Outsourced Frequent Itemset Mining in Data-Mining-As-a-Service Paradigm." IEEE Transactions on Services Computing 9.1 (2016): 18-32.

[7] http://www.ueltschi.org/teaching/chapShannon.pdf[online]

[8] Ravi, Kumar, and Vadlamani Ravi. "A novel automatic satire and irony detection using ensembled feature selection and data mining." Knowledge-Based Systems 120 (2017): 15-33.

[9] Koh, Yun Sing, and Sri Devi Ravana. "Unsupervised Rare Pattern Mining: A Survey." ACM Transactions on Knowledge Discovery from Data (TKDD) 10.4 (2016): 45.

[10] Yuan, Xiuli. "An improved Apriori algorithm for mining association rules." AIP Conference Proceedings. Vol. 1820. No. 1. AIP Publishing, 2017.

[11] He, Pinjia, et al. "An evaluation study on log parsing and its use in log mining." Dependable Systems and Networks (DSN), 2016 46th Annual IEEE/IFIP International Conference on. IEEE, 2016.

[12] Jing, Liu. "Clustering Algorithm in Data Mining Based on Web Log." International Journal of Simulation--Systems, Science & Technology 17.36 (2016).

[13] Dharmarajan, K., and Dr MA Dorairangaswamy. "Web Usage Mining: Improve The User Navigation Pattern Using Fp-Growth Algorithm." Elysium journal of engineering research and management (EJERM) 3.4 (2016).

[14] Wang, Peng, Xikun Ma, and Jingjie Yu. "An Effective Network Security Log Mining Algorithm based on Fuzzy Clustering." Applied Mathematics & Information Sciences 10.1 (2016): 307.

[15] Fernandez, F. Mary Harin, and R. Ponnusamy. "Data preprocessing and cleansing in web log on ontology for enhanced decision making." Indian Journal of Science and Technology 9.10 (2016).

[16] Mahendra Pratap Yadav ,Pankaj Kumar Keserwani and Shefalika Ghosh Samaddar ." An Efficient Web Mining Algorithm for Web Log Analysis: E-Web Miner " ,1st Int'l Conf. on Recent Advances in Information Technology | RAIT-2012