International Journal of scientific research and management (IJSRM)
||Volume||4||Issue||06||Pages||4338-4346||2016||
Website: www.ijsrm.in ISSN (e): 2321-3418

"Modified Apriori Algorithm For Efficient Web Navigation Data Mining"

Mr. Pradeep Kumar Shriwas, Mr.Deepesh Dewangan

(M.Tech. Object Oriented Software Development)

(Department of Computer Science: Kalinga University, Raipur)

shriwasji.ps@gmail.com

(Assistant Professor)

(Department of Computer Science: Kalinga University, Raipur)

deepesh.dewangan@gmail.com

Abstract: Web Data mining can be defined as knowledge discovery and analysis of useful information from the web. It is the process of determining user accessibility pattern, during the mining of log files and associated data from a particular Web site. Accurate web log mining results and efficient online navigational pattern prediction are undeniably crucial for tuning up websites and consequently helping in visitor's retention. Like any other data mining task, web log mining starts with data cleaning and preparation and it ends up discovering some hidden knowledge which cannot be extracted using conventional methods. We are proposing an enhancement to the web log mining process and to the online navigational pattern prediction named Dynamic Apriori algorithm for mining both frequent and closed frequent item set over data stream in log file. The algorithm is appropriate for noticing latest or new changes in the set of frequent item set by modifying the traditional Apriori algorithm and Travelling Salesmen Problem (TSP) in the arriving data stream.

The proposed modified algorithm we avoid unwanted and repetition of data. Here experimental analysis also use for improvement of web design, efficiently finding the User web navigation pattern which will help to grow the customer satisfaction and analysis.

Keyword: Apriori Algorithm, web navigation, data mining, knowledge discovery, web log files TSP.

INTRODUCTION

WEB DATA MINING

The term Web Data Mining is a technique used to crawl through various web resources to collect required information, which enables an individual or a company to promote business, understanding marketing dynamics, new promotions floating on the Internet, etc. There is a growing trend among companies, organizations and individuals alike to gather information through web data mining to utilize that information in their best interest.

Data Mining is done through various types of data mining software. These can be simple data mining software or highly specific for detailed and extensive tasks that will be sifting through more information to pick out finer bits of information. For example, if a company is looking for information on doctors including their emails, fax, telephone, location, etc., this information can be mined through one of these data mining

software programs. This information collection through data mining has allowed companies to make thousands and thousands of dollars in revenues by being able to better use the internet to gain business intelligence that helps companies make vital business decisions.

WEB USAGE MINING

Web usage mining is the third category in web mining. This type of web mining allows for the collection of Web access information for Web pages. This usage data provides the paths leading to accessed Web pages. This information is often gathered automatically into access logs via the Web server. CGI scripts offer other useful information such as referrer logs, user subscription information and survey logs. This category is important to the overall use of data mining for companies and their internet/ intranet based applications and information access.

WEB LOG

A weblog, sometimes written as web log or Weblog is a Web site that consists of a series of entries arranged in reverse chronological order, often updated on frequently with new information about particular topics. The information can be written by the site owner, gleaned from other Web sites or other sources, or contributed by users. Weblog is the name of a software product from South Korea that analyzes a Web site's access log and reports the number of visitors, views, hits, most frequently visited pages, and so forth.

The Scope of Data Mining

- **Automated prediction of trends and behaviours**. Data mining automates the process of finding predictive information in large databases. Questions that traditionally required extensive hands-on analysis can now be answered directly from the data — quickly.
- **Automated discovery of previously unknown patterns**. Data mining tools sweep through databases and identify previously hidden patterns in one step.

Data Mining Applications

Data mining is highly useful in the following domains −

- Market Analysis and Management
- Corporate Analysis & Risk Management
- Fraud Detection

Mining of Frequent Patterns

Frequent patterns are those patterns that occur frequently in transactional data. Here is the list of kind of frequent patterns −

- **Frequent Item Set** − It refers to a set of items that frequently appear together, for example, milk and bread.
- **Frequent Subsequence** − A sequence of patterns that occur frequently such as purchasing a camera is followed by memory card.
- **Frequent Sub Structure** − Substructure refers to different structural forms, such as graphs, trees, or lattices, which may be combined with item−sets or sub sequences.

Data Mining - Knowledge Discovery

Some people don't differentiate data mining from knowledge discovery while others view data mining as an essential step in the process of knowledge discovery. Here is the list of steps involved in the knowledge discovery process –

Data Cleaning − in this step, the noise and inconsistent data is removed. **Data Integration** − in this step, multiple data sources are combined. **Data Selection** − in this step, data relevant to the analysis task are retrieved from the database. **Data Transformation** − in this step, data is transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations. **Data Mining** − in this step, intelligent methods are applied in order to extract data patterns. **Pattern Evaluation** − in this step, data patterns are evaluated. **Knowledge Presentation** − in this step, knowledge is represented.

RELATED WORK

In the terms of data mining and pattern navigation there is lots of works has done like Refined time-out based heuristic for session identification. Which suggesting the usage of a specific density based algorithm for navigational pattern discovery [1] .Markov Model OF Web Mining consists of three different categories, namely Web Content Mining, Web Structure Mining, and The approaches result in prediction of popular web page or stage and user navigation behaviour[2]. algorithm named *Variable- Moment* for mining both frequent and closed frequent item set over data stream[3]Based on the range, a data mining method can be able to identify when a transactions becomes stale or needs to be disregarded. The bonding of memory and time usage is compared by means of Apriori algorithm and improved Frequent Pattern Tree algorithm [4.Ashish Gupta].unique approach for ranking and adaptation of Web Services using Association Rule Mining based on our proposed Semantic Logs and Semantic extension of FP-Growth [5]

Web usage mining is a special area of web mining which is based upon the discovery and analysis of web usage patterns from web logs so as to effectively and efficiently serve the needs of the users visiting the websites. [6]This information can then be used by the website administrators for efficient administration and personalization of their websites and thus the specific needs of specific communities of users can be fulfilled and so the profit can be increased. Also analyse the dynamic apriori algorithm for the mining data by the association rules.

PROBLEM INDENTIFICATION

Problem in web Uses Mining

Web Usage Mining is defined as the process of applying data mining techniques to the discovery of usage patterns from Web logs data which to identify Web user's behaviour . Web Usage Mining is the type of Web mining activity that involves an automatic discovery of user access patterns from one or more Web servers. Web Usage Mining involves determining the frequency of the page access by the clients and then finding the common traversal paths of the users. First task is the data is collected from web server log file.

A. Data Pre-Processing

Web prediction models are generally used to identify the most probable future action for sequence of requests for the following class of users:

1. Specific users (client based models are more suitable)
2. Similar minded users (e.g. group of students in the same research team)

3. General users (e.g. for wide range of users in an internet cafe)

B. Pattern Discovery & Pattern Analysis

Pattern Discovery Tools implement techniques from data mining, psychology, and information theory on the Web traffic data collected. To discover the novel, potentially useful and interesting information, several methods and data mining algorithms are applied such as Path Analysis, Association Rules, Sequential Patterns, Clustering, and Classification.

3.1 Problem Identification in Apriori Algorithm of mining association rules:

1. Assume transaction database is memory resident.
2. Require many database scans.
3. Apriori, while historically significant, suffers from a number of inefficiencies or trade-offs, which have spawned other algorithms. Candidate generation generates large numbers of subsets (the algorithm attempts to load up the candidate set with as many as possible before each scan). Bottom-up subset exploration (essentially a breadth-first traversal of the subset lattice) finds any maximal subset S only after all $2^{|S|} - 1$ of its proper subsets.
4. Problem related to pattern prediction.
5. Problem related to log cleaning.
6. Problem related to log acquisition of data.

APRIORI ALGORITHM

Procedure **Apriori** (T, *minSupport*)

{

//T is the database and *minSupport* is the minimum support

L1= {frequent items};

for (k= 2; L_{k-1} != Ø ; k++) **{**

C_k= candidates generated from L_{k-1}

//that is Cartesian product L_{k-1} x L_{k-1} and eliminating any k-1 size item set that is not //frequent

for each transaction **t** in database **do{**

#increment the count of all candidates in C_k that are contained in t

L_k = candidates in C_k with *minSupport*

}//end for each }//end for

Return $\mathbf{U}_k\mathbf{L}_k$;

}

DYNAMIC APRIORI ALGORITHM

INPUT: Transactional database No. of different Items.

METHOD:

1. Scan each transaction in database and generate only one nth candidate Itemset based on size transaction.

    1.1 Merge candidate nth itemset based on size of transaction and if any itemset is repeated increment the support count of that item set.
    1.2 Compare support count with minimum support count.
    1.3 If support count is minimum then min support add it to the frequent n item set.
    1.4 From the result of previous step use subset property and generate all frequent set (if itemset – k size is frequent then all its subset are frequent)

2. Using Vertical data format find all the possible frequent itemset(for those which are not considered as frequent in step 1)

    1.1 if support count of item set is greater then 0 then check result in step1. If itemset is present in step1 result then considered frequent also in step 2 . and if it is not present in the result of step 1 then delete it.

2. Merge the result of step1 and step 2.

    2.1 if the same item set is appear two times then write it once and add their support count in the final result.

3. All Possible frequent Itemset.

PRAPOSED ALGORITHM

Association rules are if/then statements that help uncover relationship between seemingly unrelated data in a relational database or other information repository.

The target is to find pages that are accessed together by majority of the user and hence should be linked in a proper way in order to maximize user satisfaction by providing to the user the access flow they expect.

Step 1: Original data.

Step 2: Cleaning through regular expression matching algorithm.

Step 3: Classification of potential user through thresh holding method.

Step 4: Clustering of user navigation prediction through weighted graph algorithm.

Step 5: Using association rule (Frequent pattern) for Prediction Engine1.

Step 6: Using LCS algorithm for prediction engine2.

Step 7: for decision making compare and take out the common of Step 5 & 6.

EXPERIMENTAL TOOL FOR PRAPOSED WORK

MATLAB –R2013a

MATLAB is a high-performance language for technical computing. It integrates computation, visualization, and programming in an easy-to-use environment where problems and solutions are expressed in familiar mathematical notation. Typical uses include:

- Math and computation
- Algorithm development
- Modelling, simulation, and prototyping
- Data analysis, exploration, and visualization
- Scientific and engineering graphics
- Application development, including Graphical User Interface building

The MATLAB System: The MATLAB system consists of five main parts:

1. The MATLAB language. 2. Handle Graphics. 3. The MATLAB mathematical function library. 4. The MATLAB working environment. 5. The MATLAB Application Program Interface (API).

MATLAB is the easiest and most productive software for engineers and scientists. Whether you're analyzing data, developing algorithms, or creating models, MATLAB provides an environment that invites exploration and discovery. It combines a high-level language with a desktop environment tuned for iterative engineering and scientific workflows.

RESULT & DISCUSSION

The system should be able to provide best prediction for web navigation patterns.

The performance will be measured by:

- Minimum support
- Mean Length
- Browsing session

After applying the frequent pattern algorithm. The next step is to get LC Pattern from the rules generated by the Association algorithm. Theses longest pattern are generated through LCS algorithm.

PERFORMANCE EVALUATION

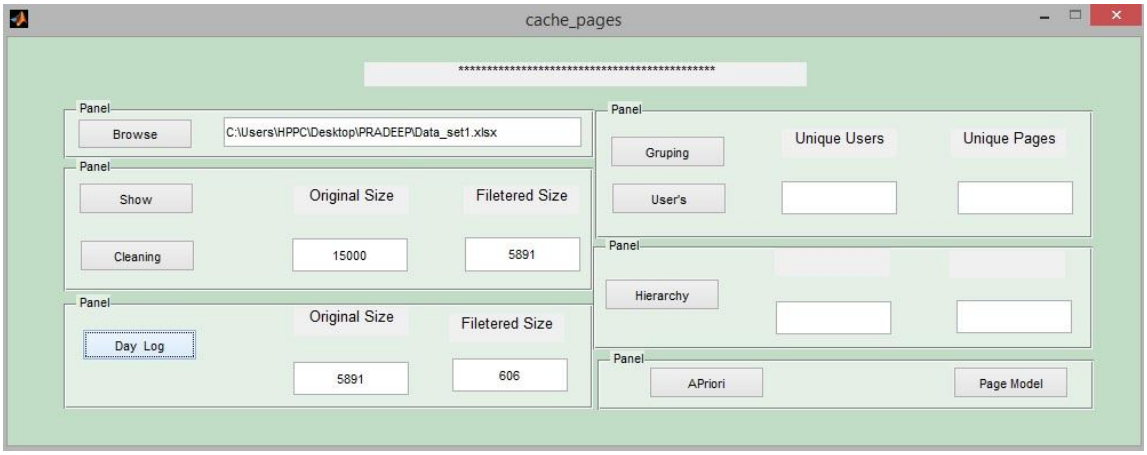

Figure 10: Loading data set cleaning and Separating Day Log

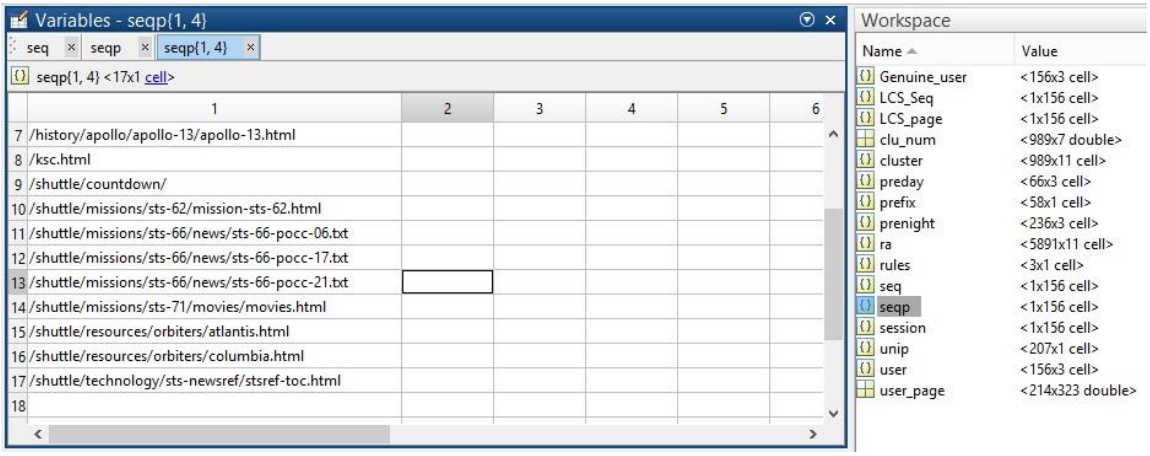

Table 10: Generated Frequent Pattern

Here we compare our proposed system with Apriori, More and Graph Traverse algorithm. We change the minimum support threshold and take the experimental results which are shown in following figure. While the minimum support becomes lower, the number of frequent patterns increases and the frequent patterns get longer.

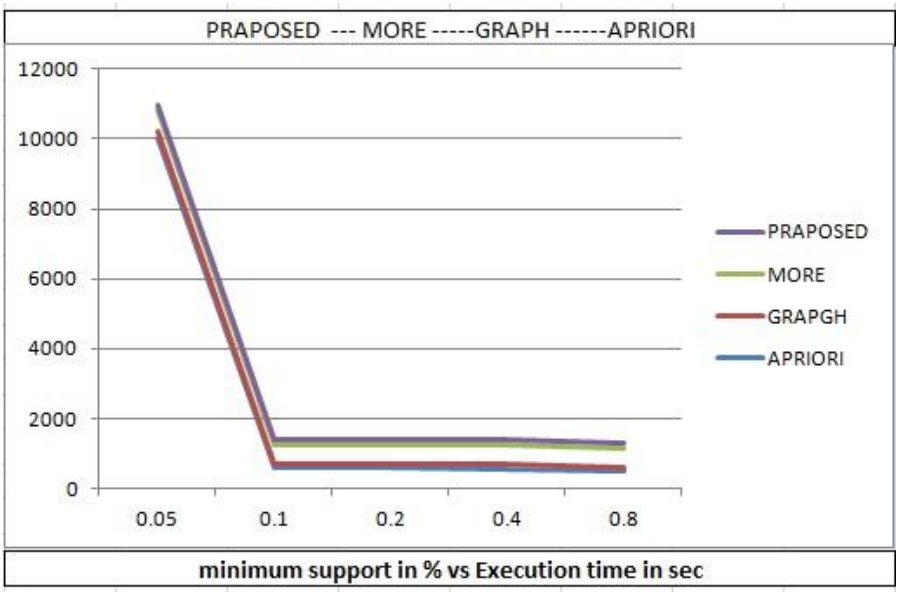

Figure 12 : Minimum support vs Execution time

Scalability of the algorithms by varying mean length and number of web browsing session is illustrated here.

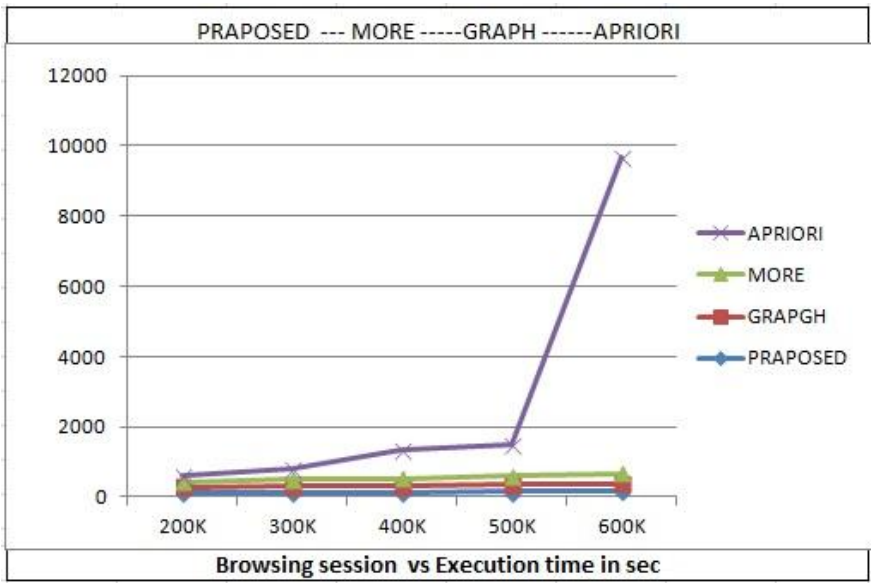

Figure 13: Browsing session vs Execution time

Actually the execution time is directly proportional to number of browsing sessions. As the length of web browsing session or dataset increases, the cost of scanning dataset also increase. Same scenario is confirmed in experimental result.

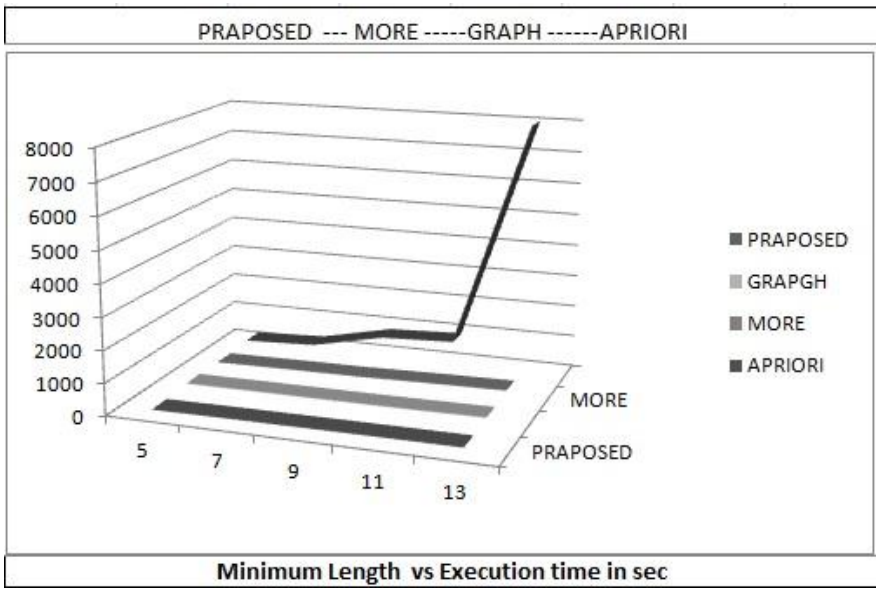

Figure14: Minimum length vs Execution time

CONCLUSION

Our proposed system objectives is using Web mining technology to improve web services, to improve web design, customer satisfaction and efficient market analysis and minimizing execution time. Here we modify

the Apriori algorithm and with the use of association rule work to mine the web data and finding the pattern to navigate the website. Our work helped to improve efficiency in terms of browsing session, mean length and minimum support. We have also compared the performance with another algorithm like apriori, graph traverse etc the resultant of evaluation is completely satisfying the user need regarding web navigation.

SCOPE OF FUTURE WORK

The future scope of these technology is that we can work for the Big Data evaluation in future, It can be used in cache coherence, It Can be work in Clustering algorithm It can also work with the other Algorithms used in web navigation or web uses mining.

REFERENCES

[1] Show-Jane Yen*, Yue-Shi Lee* and Min-Chi Hsieh, "An Efficient Incremental Algorithm for Mining Web Traversal Patterns", Proceedings of the 2005 IEEE International Conference on eBusiness Engineering (lCEBE'05)

[2] Ming-Syan, Jong Soo, Philips Yu., "Efficient Data Mining for Path Traversal Patterns", ProceedingOf 1998 Knowledge and Data Engineering, IEEE on (Volume: 10, Issue: 2)

[3] Ming-Yen Lin and Suh-Yin Lee, "Incremental Update on Sequential Patterns in Large Databases", Tools with Artificial Intelligence, 1998. Proceeding of Tenth IEEE International Conference.

[4] Jian Pei, Jiawei Han, Behzad and Hua Zua "Mining Access Patterns Efficiently from Web Logs", Proceeding in 2000 on National Sciences and Engineering Research Council of Canada, Hewlett-Packard Lab.

[5] R. Sri kant, Rakesh Agrawal, "Mining Sequential Patterns: Generalization and Performance Improvements", IBM AJmaden Research Center (1996).

[6] Show-Jane Yen and Arbee L.P. Chen "A Graph-Based Approach for Discovering Various Types of Association Rules", IEEE Transactions on Knowledge and Data Engineering, Vol. 13, No. 5, September/October 2001.

[7] Anna Gutowska and Luis L. Perez "A Comparison of Methods for Classification and Prediction of Web Access Patterns", COMP540 - Final Report - Spring 20 I O.

[8] Arthur.A.Shaw, N.P. Gopalan "Frequent Pattern Mining of Trajectory Coordinates using Apriori Algorithm", proceeding in International Journal of Comp. Applications in volume 22-9, May 2011.

[9] Introduction to data mining with case studies, G. K. Gupta, PHI 2006.

[10] More S. "Modified Path Traversal for an Efficient Web Navigation Mining" Proceedings of the 2014 IEEE International Conference on Advanced Communication Control and Computing Technologies (ICACCCT).

[11] Bhargav A., Bhargav M. "Pattern Discovery and Users Classification Through Web Usage Mining" Proceedings of the 2014 IEEE International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT).

[12] Ying, Jia-Ching, Chin, Chu-Yu, Tseng, Vincent S. "Mining Web Navigation Patterns with Dynamic for Navigation Prediction" Proceedings of the 2012 IEEE International Conference on Granular Computing (GrC).