INTERNATIONAL JOURNAL OF TECHNOLOGY AND COMPUTING (IJTC)
ISSN-2455-099X,
Volume 2, Issue 9 September 2016

Web Usage Mining Through FP Split and APRIORI Algorithm

Ms. Manjot Kaur[1], Ms. Rupinder Kaur Gurm[2]

[1]Research Scholar, [2]Associate Professor

[12]Department of CSE, RIMT-IET

[1]manjotnagpal89@gmail.com, [2]rupindergurm@gmail.com

Abstract: In the recent years the explosive growth of WWW has turned the web into the largest source for available online data. Situations such as several unrelated topics in a single web page may lead to confusion and make it harder to reach to the information that the visitors are looking for. The design of whole website (structure, usability, interface, content, etc.) is one of the most important aspects for an institution that want to survive in the cyberspace. Understanding the way users browse the website and finding out which of them is the most frequently used link and pattern for using the features available in the site is also a must. It is a novel technique to use APRIORI algorithm for the same (i.e. weblog/web usage mining). All this information is available online but is hidden from the users. This research work uses web usage mining (WUM) Apriori and FP Growth based approach for analyzing the visitor's browsing behaviour.

Indexed Terms: **APRIORI, WWW, Web Mining, Cyberspace, Mining**

I. INTRODUCTION

Web is a vast, explosive, diverse, dynamic and mostly unstructured data repository, which supplies an incredible amount of information, and also increases the complexity of dealing with the info from different perspectives of users, web service providers, view, business analysts. The users would desire to have effective search tools to find relevant information precisely as well as easily. The Web service providers want to find the best way to predict the users' behaviours and personalize information to reduce the traffic load and design the Web site that is suited for different group of users. The business analysts also want to have tools to learn the consumer/users' wants and needs. All of them are expecting tools or techniques to help them satisfy their demands and/or solve the problems they encounter on the Web. Therefore, Web mining becomes a popular and active area and is taken as the research topic for this investigation.

The amount of data kept in databases on the internet is increasing at a rapid pace due to which World Wide Web has become a dominant medium to retrieve information and mine knowledge. But due to the following characteristics of Web data and the below mentioned problems faced by users, they find it difficult to fully utilize the facts and find significant information accessible on the Web:

A. Characteristics of Web data:

Massive amount of data: The data on the Web is enormously large in size. This makes it hard to implement traditional data mining techniques on the data.

Distributed: The data on the Web is dispersed across computers spread over diverse locations. So before mining there is a need to gather the documents together.

Heterogeneous: Web data comprises of varied types of data such as textual data as well as multimedia data.

Semi structured: Web documents have a structure that allows for querying but that is without a query language. Hence to process the documents easily they should be represented in a correct format.

Dynamic: The content and the structure of Web documents keeps on changing. Hence keeping up with the change is essential so that information retrieval is not affected.

Web Usage Mining is the application of data mining techniques to discover exciting usage patterns from Web data, to understand and serve the requirements of Web-based applications in a better manner. Usage data captures the identity or origin of the Web users along with their browsing behavior at a Web site. Itself web usage mining can be classified further depending on the kind of usage data being considered. There are web server data, application server data and application level data. Web server data corresponds to the user logs that are collected at the web server. Some of the typical data collected there include IP addresses, access time and page references of the users. This work focuses on web usage mining and in particular concentrates on discovering the web usage patterns of websites from the server log files.

B. Problems related to Web Mining:

Discovering the appropriate information: Internet users usually find it difficult to find the useful information due to low precision and low recall.

Finding novel knowledge from existing information: This problem includes extracting novel patterns from the data, but then to accomplish this task one needs the accessibility of repository of Web data.

Information personalization: Due to the diversity in the preferences of the users, they desire different contents and presentations of information. Hence it imposes a stress on the part of Web developer/designer to present custom-made Web pages to the user.

The following steps are included in web usage mining:

1. **Data Collection**
2. **Preprocessing**
   2.1. Data cleaning
   2.2. User identification

2.3. Session identification
2.4. Path completion

3. Knowledge discovery

4. Pattern analysis

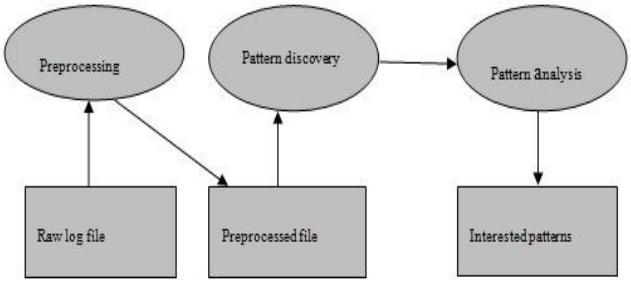


Figure: The complete process of web usage mining

II. RELATED WORK

Pooja Sharma and Rupali Bhartiya in their research paper **An Efficient Algorithm for Improved Web Usage Mining [1]** shared that this paper presents a new algo to discover data clusters for numerical and nominal data. Apriori algorithm generates quite a large number of candidate set which is not efficient for both data.

Initial research of **1994**, **Fast Algorithms for Mining Association Rules** by **Rakesh Agrawal and Ramakrishnan Srikant [2]**shows how only the best features of the two proposed algorithms can be combined into a hybrid algorithm which they named as AprioriHybrid. Experiments show that AprioriHybrid scales linearly with the no. of transactions and it also has excellent scale-up properties w.r.t. the no. of items in the database and the transaction size.

Sandeep Singh Rawat and Lakshmi Rajamani in **August 2010** in their research paper **Discovering Potential User Browsing Behaviors Using Custom-Built Apriori Algorithm [3]** mentioned that Web Usage Mining is collecting the interesting patterns using the required interestingness measures, which in turn help the designer in discovering the sophisticated patterns that would be used for developing the website. It proposed a custom-built apriori algorithm to find the effective pattern analysis.

Drawbacks:

1. adoption of more up-to-date techniques/tools such as fuzzy association rule mining algo

2. testing of the work that has been proposed on a real distributed platform.

3. use of grid-computing paradigm in order to solve more challenging web mining issues

Another research of **2010**, **An Algorithm for Frequent Pattern Mining Based On Apriori [4]** by **Goswami D.N., Chaturvedi Anshu, Raghuvanshi C.S.** stated that with passing time a number of changes have been proposed in Apriori to enhance its performance in terms of time and number of database passes. Here, three different frequent pattern mining techniques (Record filter, Intersection and Proposed Algorithm) are given which are based on classical Apriori algorithm. Out of these approaches Record filter approach has proved better than classical Apriori Algorithm, Intersection approach has proved to be better than Record filter approach and finally the proposed algo proved that it is better than other frequent pattern mining algo.

Ankit R Kharwar, Viral Kapadia, Nilesh Prajapati and Premal Patel in **May 2011** in their research paper **Implementing APRIORI Algorithm onWeb server log [5]** shared that in order to produce the usage patterns and user behaviors, this paper implements the process of Web Usage Mining using basic association rules algo called Apriori algo. Lastly, this paper presented finding association rule from server log which are useful in many applications such as cache for web page, targeted advertising, marketing etc.

Drawback: The output of the system was only in terms of memory usage and the speed of producing association rules leaving out many other important aspects.

Another research of the year **September 2011** titled **Web Log Mining using Improved Version of Apriori Algorithm [6]** by **Suneetha K R and Krishnamoorti R** stated that a novel method of top down approach is proposed in order to reduce repetitive disk read. The improvised version of Apriori Algo greatly reduces database scans and avoids the generation of unnecessary patterns reducing database scans, time and space consumption.

According to **R. Suguna and D. Sharmila** in their paper **An Overview of Web Usage Mining [7]** published in **February 2012,** mentioned that many researchers have already done a huge variety of work on web content mining and web usage mining in order to improve the efficiency of the websites. They do so by providing novel methods and this paper gives an overview about of the existing works done by those researchers on WUM.

Another research of the year **2012** titled **Discovery of Frequent Patterns from Web Log Data by using FP-Growth algorithm for Web Usage Mining [8]** by **Mr. Rahul Mishra and Ms. Abha Choubey** stated that in this paper they used the FP-growth algo for obtaining frequent access patterns from the web log data and providing the valuable information about the user's interest.

Drawback: Their experimental result shows that the FP-growth method is efficient and scalable for mining both long and short frequent patterns but the algo could have also been extended to web content mining and web structure mining.

Another research, **A Study: Web Data Mining Challenges and Application for Information Extraction [9]** by **T. Sunil Kumar, Dr. K. Suvarchala** that was published in **November-December 2012**stated that the aim of their paper was to explore the role of data mining for information extraction in web content, structure and usages mining in current web models, and outlines the process of extracting patterns from the data. Issues regarding how to integrate a data mining system with a database were also discussed in this paper.

Drawback: This approach generates quality recommendations by evaluating collective efforts rather

than basing recommendations on only single person's past experience.

According to **Bina Kotiyal, Ankit Kumar, Bhaskar Pant, R.H. Goudar, Shivali Chauhan and Sonam Junee**as stated in their paper **User Behavior Analysis in Web Log through Comparative Study of Eclat and Apriori [10]** published in **December 2012,** this paper is based on two intelligent algorithms for predicting the user behaviors- Apriori and Eclat and also does the performance comparison of the two algorithms in terms of time and space complexity for the filtered data.

Drawback: Both the algorithms helps to find out the navigation behavior of the user based on the previous visits although it is very much clear that Eclat algorithm serves better for the large databases as it generates less tables and therefore less time it takes to perform the analysis.

Kirti S. Patil and Sandip S. Patil in **Jan 2013** in their paper **Sequential Pattern Mining Using Apriori Algorithm & Frequent Pattern Tree Algorithm [11]**stated that a new algorithm is proposed is this paper; it combines the Apriori and FP-tree structure which is proposed in FP-growth algorithm. The advantage of the proposed algo is that it does not need to generate conditional pattern bases and sub- conditional pattern tree recursively.

Drawback: It needs to be optimized for counting the support of the candidates and expanded for mining more larger database.

Latheefa.V and Rohini.V in their research titled **Web Mining Patterns Discovery and Analysis Using Custom-Built Apriori Algorithm [12]**of**March 2013** presents the idea for a custom-built Apriori Algo for the discovery of frequent patterns in the web log data. This research also includes development of a tool to discover the frequent patterns, association rules in the web log data. The experiments that have been conducted in this study proved that custom-built Apriori Algo is efficient than Classical Apriori Algo as it takes lesser time.

Drawback: This analysis should be extended by implementing other interesting measures such as lift in order to fine tune the results.

Yet another research of **2013** titled **A Modified FP-Tree Algorithm for Generating Frequent Access Patterns [13]** by **Harendra Singh, Ashish Kumar Srivastava and Sitendra Tamrakar** stated that the algorithm that is proposed is not generating any candidate sets, rather larger number of patterns would be generated. Because of which the number of tree traversals would be more. The obtained result shows that the proposed algorithm in all instances takes 25% lesser time compared to the Classic Apriori algorithm.

Lastly, **Shipra Khare, Prof Vivek Jain and Prof Manoj Ramaiya** in their research titled **Implementation of Web Usage Mining with Customized Web Log Using FP Growth Algorithms [14]**carried out in**September 2013** stated that this paper implements the process of Web Usage Mining using basic association rules algorithm called as Apriori Algorithm. This piece of research work concentrates on web usage mining and in particular focuses on discovering the web usage patterns of websites from the server log files. The comparison of time and memory usage is compared using Apriori algorithm and Frequent Pattern Growth algorithm.

III. PROBLEM FORMULATION

In this research, we will emphasize on Web usage mining. Reasons for this are very simple: With the E-commerce explosion, the way companies are doing businesses has changed. E-commerce, mainly characterized by electronic transactions through Internet, has provided us an effective way and a cost-efficient of doing business.

The E-businesses growth is astonishing, considering how e-commerce has made flipkart.com become the so-called “on-line one & all store”. Unluckily, for most companies, web is nothing more than a marketplace where transactions takes place. They do not realize that millions of visitors interact daily with their web sites around the globe, hence massive amounts of data is being generated. And also they did not realize that this info could be very worthy to the company in the fields such as improving customer services and relationship, understanding customer behaviour, measuring the success of marketing efforts, launching target marketing campaigns, and the list continues. Uses of web usage mining:

- Enhanced performance of the server
- Improvised web site navigation
- Improvised web application’s system design
- Customers targeted for e-commerce
- Identify prime advertising locations

IV. PROPOSED WORK

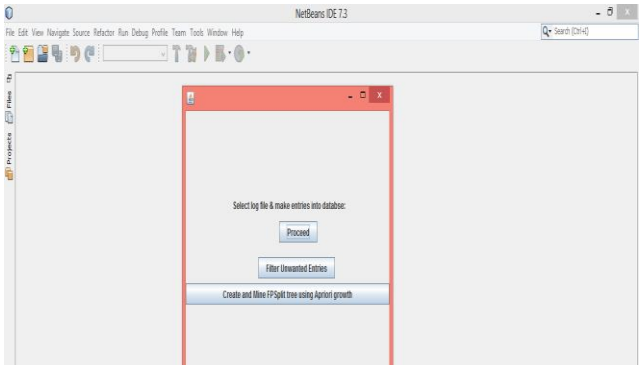
The main goal of the proposed system is to identify usage pattern from web log files of a website. Apriori and FP Growth Algorithm are mostly used for this purpose. Both of these are influential algorithms for mining frequent item sets for boolean association rules.

In spite of being simple, Apriori algorithm suffers from certain limitations such as cost to handle a huge number of candidate sets and the effort to repeatedly scan the database and check a large set of candidates by pattern matching, which is true for mining long patterns especially. Using Apriori algorithm for web usage mining is therefore not efficient.

In order to overcome the drawback inherited in Apriori, an efficient FP-tree based mining method, FP-growth is used, which contains two phases, where an FP tree is constructed in the first phase, and the second phase recursively searches the FP tree and outputs all frequent patterns. The main drawback of FP-growth algorithm is the explosive lack of a good candidate generation method.

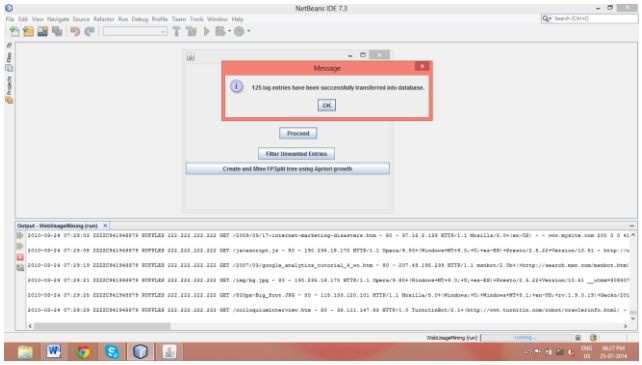
V. EXECUTION AND EXPERIMENTAL RESULTS

Choosing the Log file for web usage mining.

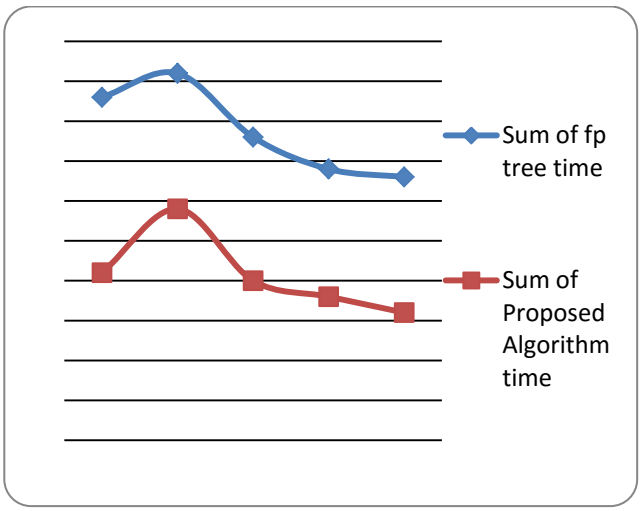


The above screenshot is for the first screen where we first transfer the contents of log file into the database and then do its cleaning to remove the records where where some image, xml, css, js or some other support file is being requested rather the main webpage. This is called filtering the log.

Filtering the unwanted entries

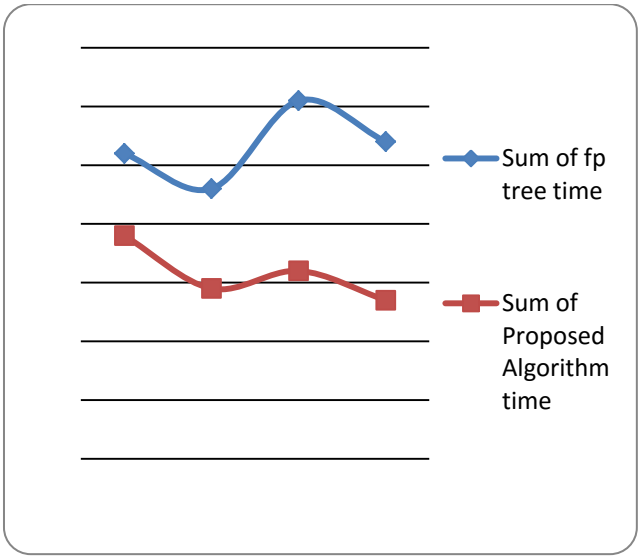


Comparison of proposed algorithm with fp tree.



No. of records	FP tree time	Proposed Algorithm Time
127	46	29
116	43	21
112	38	20
101	34	18
92	33	16

The above graph shows the comparison between FP Tree and FP Split tree methods for support count 3. We can see that the time taken by proposed algorithm is always better than the traditional method. The difference would be more clearly visible if we get logs of few days' time for some website where the incoming traffic is also more frequent. The proposed algorithm has been tested on logs received from Kurukshetra University website for only 29 minutes. The more is the no. of records, the better visible is the difference between the efficiency of the two algorithms.



Support	FP tree time	Proposed Algorithm Time
2	52	38
3	46	29
4	61	32
5	54	27

The above graph shows the comparison between FP Tree and FP Split tree methods for different support counts for a fixed no. of records.

VI. CONCLUSION AND FUTURE WORK

The main drawback of the all-time classic, Apriori algorithm, is that the candidate set generation is very costly, especially if for a large number of patterns and/or long patterns. The drawback of FP-growth algorithm is the lack of a good candidate generation method. Therefore we combined the benefits of both by using candidate generation of Apriori algorithm and then feedings its result into the FP Tree to reduce its cost and continue with the FP Growth algorithm hence swiping out all the limits by our newly generated algorithm.

Future research could be directed towards finding a more efficient algorithm than the one proposed above. In future

the algorithm could also be extended to web structure mining, web content mining, etc. This work could also be extended to extract information from the image files.

7. REFERENCES

[1] Pooja Sharma, Rupali Bhartiya, "*An Efficient Algorithm for Improved Web Usage Mining*", International Journal of Computer Technology &Applications || Vol. 3 (2), 766-769 || ISSN:2229-6093

[2] Rakesh Agrawal, Ramakrishnan Srikant, "*Fast Algorithms for Mining Association Rules*", Proceedings of the 20th VLDB Conference, Santiago, Chile, 1994

[3] Sandeep Singh Rawat, Lakshmi Rajamani, "*Discovering Potential User Browsing Behaviors Using Custom-Built Apriori Algorithm*", International Journal of Computer Science & Information Technology (IJCSIT) || Vol.2, No.4, August 2010

[4] Goswami D.N., Chaturvedi Anshu, Raghuvanshi C.S., "*An Algorithm for Frequent Pattern Mining Based On Apriori*", (IJCSE) International Journal on Computer Science and Engineering || Vol. 02, No. 04, 2010 || 942-947

[5] Ankit R Kharwar, Viral Kapadia, Nilesh Prajapati, Premal Patel, "*Implementing APRIORI Algorithm on Web server log",* National Conference on Recent Trends in Engineering & Technology, 13-14 May 2011

[6] Suneetha K R, Krishnamoorti R, "*Web Log Mining using Improved Version of Apriori Algorithm*", International Journal of Computer Applications || Volume 29– No.6, September 2011 || ISBN: 0975 – 8887

[7] R. Suguna, D. Sharmila, "*An Overview of Web Usage Mining*", International Journal of Computer Applications || (0975 – 8887) || Volume 39– No.13, February 2012

[8] 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 || Volume 2, Issue 9, September 2012 || ISSN: 2277 128X

[9] T. Sunil Kumar, Dr. K. Suvarchala, "*A Study: Web Data Mining Challenges and Application for Information Extraction*", IOSR Journal of Computer Engineering (IOSRJCE) || ISSN: 2278-0661, ISBN: 2278-8727 || Volume 7, Issue 3 (Nov. - Dec. 2012) || PP 24-29

[10] Bina Kotiyal, Ankit Kumar, Bhaskar Pant, R.H. Goudar, Shivali Chauhan and Sonam Junee, "*User Behavior Analysis in Web Log through Comparative Study of Eclat and Apriori*", Proceedings of 7th International Conference on Intelligent Systems and Control (ISCO 2013) || 978-1-4673-4603-0

[11] Kirti S. Patil, Sandip S. Patil, "*Sequential Pattern Mining Using Apriori Algorithm &Frequent Pattern Tree Algorithm*", IOSR Journal of Engineering (IOSRJEN) || e-ISSN: 2250-3021, p-ISSN: 2278-8719 || Vol. 3, Issue 1 (Jan. 2013), ||V4|| PP 26-30

[12] Latheefa.V, Rohini.V, "*Web Mining Patterns Discovery and Analysis Using Custom-Built Apriori Algorithm*", International Journal of Engineering Inventions || e-ISSN: 2278-7461, p-ISSN: 2319-6491 || Volume 2, Issue 5 (March 2013) || PP: 16-21

[13] Harendra Singh, Ashish Kumar Srivastava, Sitendra Tamrakar, "*A Modified FP-Tree Algorithm for Generating Frequent Access Patterns*", JECET || June – August-2013; Vol.2.No.3 || 730-740 || E-ISSN: 2278–179X

[14] Shipra Khare, Prof Vivek Jain, Prof Manoj Ramaiya, "*Implementation of Web Usage Mining with Customized Web Log Using FP Growth Algorithms*", International Journal of Engineering & Managerial Innovations (IJEMI) || ISSN: 2321-693X || Volume I (II), September 2013