WEB USAGE MINING TO IMPROVE WEBSITE THROUGH ASSOCIATION RULE MINING USING DYNFP-GROWTH ALGORITHM

Kamalben Patel[1], Dr. Shyamal Tanna[2]

[1] Student, Information Technology, LJ Institute of Technology, Gujarat, India
[2] Assistant Professor, Information Technology, LJ Institute of Technology, Gujarat, India

ABSTRACT

We look at different techniques for successive thing set arrangements for Web Usage information like of Apriori and FP-development calculation. The successive thing set mining has been performed with the assistance of Apriori calculation. It gives us distinctive continuous thing set mining comes about with their bolster tally. Apriori peruses document once for each emphasis. Where FP-Growth is the way that the calculation just needs to peruse the record twice. In FP-Growth coming about FP-tree is not Unique for the same "legitimate" database the procedure needs two finish outputs of the database. That"s why utilize Dyn FP-Growth for finding incessant thing set .

Keyword: - *Web Usage Mining, Apriori algorithm, FP-growth algorithm, Minimum support, Association rule*

1. INTRODUCTION

Information mining is regularly characterized as the way toward finding significant, new connection examples and patterns through non-unimportant extraction of verifiable, already obscure data from substantial measure of information put away in vaults utilizing design acknowledgment and also factual and numerical methods. Information mining is essential stride of KDD (learning find from information) prepare. KDD prepare incorporate Data Selection, Data Pre-handling, Data Transformation, Data Mining, Data design assessment/translation and information representation. Information mining can be utilized to characterize information into predefined classes (arrangement), or to parcel an arrangement of examples into disjoint and homogeneous gatherings (grouping), or to recognize visit designs in the information, as conditions among ideas qualities (associations)[8] .

Information mining is characterized as the programmed extraction of obscure, helpful and justifiable examples from expansive database. Huge development of World Wide Web expands the multifaceted nature for clients to peruse viably. To build the execution of sites better site outline, web server exercises are changed according to client's interests. The capacity to know the examples of client's propensities and interests helps the operational techniques of endeavors. Different applications like etrade, personalization, site outlining, recommender frameworks are fabricated effectively by knowing clients route through web[8] .

1.1 Web Mining

Web mining is the application of data mining techniques to automatically retrieve, extract and evaluate information for knowledge discovery from web documents and services. The objects of Web mining are vast, heterogeneous and distributing documents. The logistic structure of Web is a graph structured by documents and hyperlinks, the mining results maybe on Web contents or Web structures. Web mining is divided into three types.

Web Content Mining:

Web content mining is the mining, extraction and integration of useful data, information and knowledge from Web page content. The heterogeneity and the lack of structure that permits much of the ever expanding information sources on the World Wide Web,such as hypertext documents, makes automated discovery, organization

Web Structure Mining:

Web Structure Mining mines the structure of hyperlinks within the web itself. Structure represents the graph of the link in a site or between the sites .It is the process of using graph theory to analyse the node and connection structure of a web site.

Web Usage Mining:

Web Usage Mining is the application of data mining techniques to discover interesting usage patterns from Web data in order to understand and better serve the needs of Web-based applications. Usage data captures the identity of Web users along with their browsing behaviour at a Web site.Web Usage Mining mines the log data stored in the web server.

1.2 Overview of Web Usage Mining

There are four stages in web usage mining.

Data Collection :

Users log data is collected from various sources like serverside, client side, proxy servers and so on.

Pre-processing:

Performs a series of processing of web log file covering data cleaning, user identification, session identification, path completion and transaction identification.

Pattern discovery:

Application of various data mining techniques to processed data like statistical analysis, association, clustering, pattern matching and so on.

Pattern analysis:

Once patterns were discovered from web logs, uninteresting rules are filtered out. Analysis is done using knowledge query mechanism such as SQL or data cubes to perform OLAP operations.

2. EXISTING SYSTEM

In this paper, author]discussed an approach to identify web link patterns which has been developed from web log and analysis of patterns is presented .The frequent item set mining has been performed with the help of Apriori algorithm. It gives us different frequent item set mining results with their support count. Web link sequences below support threshold are pruned. We found different frequent item set mining results by varying minimum support (2%- 3%).Association rules miner give all the possible rules with their confidence and Lift. Using the knowledge of the web site structure and the behavior of the site’s visitors, we analyzed the pruned rule set from the user ‘s point of view and proposed actions that a webmaster may decide to take based on knowledge extracted from rules in order to enhance a website and improve visitor’s browsing experience.

- Web log Extractor Module: They developed this web extractor to extract the IPs and Web links from a web log file. It gives a File 1, which is used as an input for data pre-processing module.
- Data Pre-processing Module: Data pre-processing module produces an input file for APRIORI which contains entry of different navigational profile. Here each IP works as user id and web links as item sets. They used link list data structure to make input file containing navigational profile entries for APRIORI.
- Apriori or FP Growth Algorithm Module: If a set cannot pass a test, all of its superset will fail the same test as well; it is called anti-monotone because the property is monotonic in the context of failing a test. Apriori uses above property to find L(k) from L(k-1).
- Association Rule Generation Module: This module finds out the association rules in between frequent mining pattern results. K-item set results will work as a input to association rule miner and it will give patterns as a output with their confidence value.

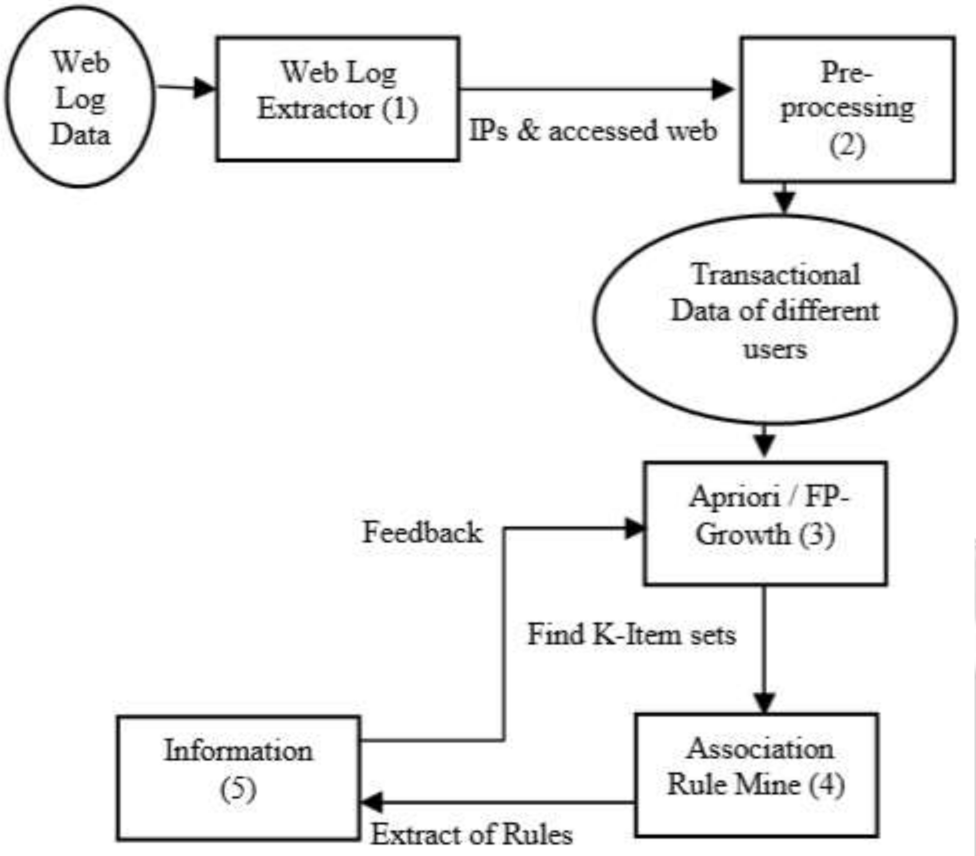

Figure 1: Proposed Model of Existing System [1]

• Information: Knowledge is derived with the help of extraction of rule those satisfy minimum confidence, below this minimum confidence other rules has been ruled out. Now knowledge may help in finding out the strongly occurring patterns

3. RESEARCH METHODOLOGY

Problem Statement: In the base paper they use Apriori algorithm, In Apriori reads file once for every iteration. Where FP-Growth is the fact that the algorithm only needs to read the file twice. Apriori has less memory usage and less runtime. FP-Growth is more scalable

3.1 Proposed System

As mentioned in the problem statement, we are going to fulfill all these demands.

Figure 2, shows the workflow of the proposed system. In this system, we proposed a DynFP-Growth Algorithm which will Generate Frequent Itemsets from web log File..

Algorithm

Step 1: Get log file

Step 2: Remove Unnecessary URLs

Step 3: Remove Robot files URLs

Step 4: Remove Log entry with status code<200 or status code>299

Step 5: User Identification Based on IP and User Agent

Step 6: Generate association rule of sequential pattern using DynFP-Growth Algorithm

Step 7: Generate sensitive and insensitive rule file as per support and confidence

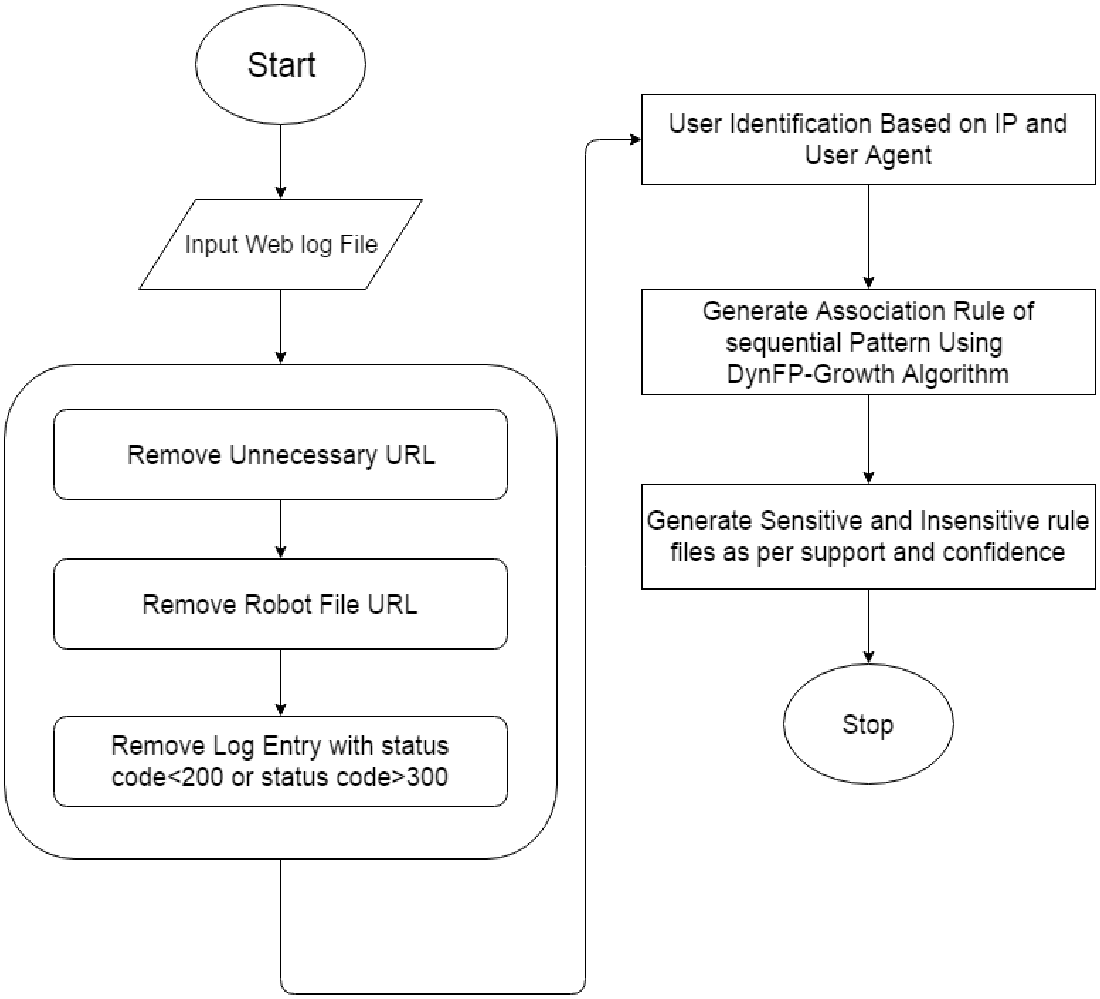
Start
Input Web log File
Remove Unnecessary URL
Remove Robot File URL
Remove Log Entry with status code<200 or status code>300
User Identification Based on IP and User Agent
Generate Association Rule of sequential Pattern Using DynFP-Growth Algorithm
Generate Sensitive and Insensitive rule files as per support and confidence
Stop

Figure 2: Proposed Work Flow

3.2 Experimental Analysis

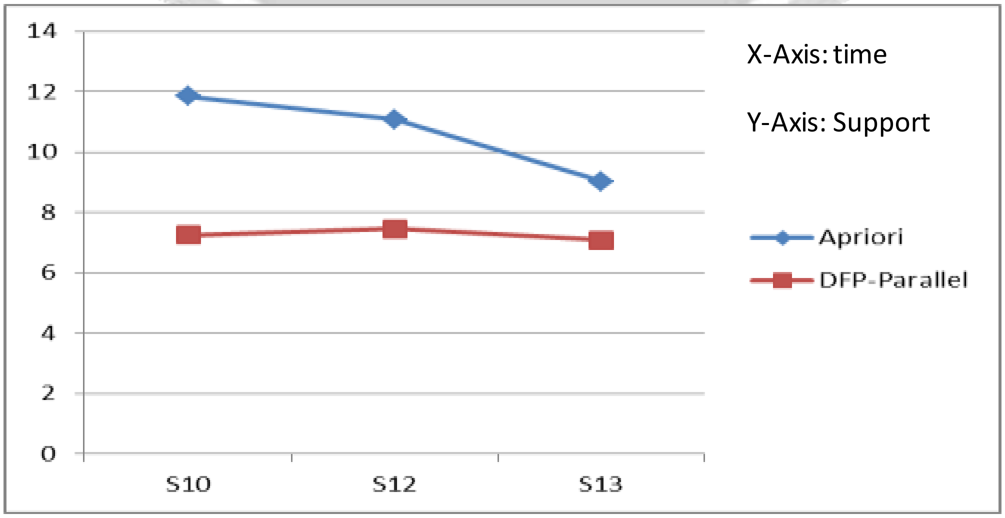

Figure 3: Comparison of Apriori with DFP-Parallel According to time

4. CONCLUSIONS

The frequent item set mining has been performed with the help of Apriori algorithm. It gives us different frequent item set mining outcomes with their help count. Apriori reads document once for every generation. Where in FP-boom is the reality that the set of rules only desires to examine the file twice. In FP-growth resulting FP-tree isn't precise for the identical “logical” database the technique wishes two entire scans of the database .That’s why i take advantage of DynFP-increase.

5. ACKNOWLEDGEMENT

I am very grateful to **Dr. A. C. Suthar,** Director of L. J. Institute of Engineering and Technology, for providing facilities to achieve the desire milestone.

I also extend my thanks to Head of Department **Prof. Gayatri Pandi** for her inspiration and continuous support.

I wish to warmly thank my guide**, Dr. Shyamal Tanna** for all his diligence, guidance, encouragement, inspiration and motivation throughout. Without his valuable advice and assistance it would not have been possible for me to attain this landmark**.** He has always been willingly present whenever I needed the slightest support from him. I would not like to miss a chance to say thank for the time that he spared for me, from his extremely busy schedule. I am very obliged to all my dear friends for their continuous livelihood and comfort in each and every phase of my life.

I would like to thank all of them whose name are not mentioned here but have played a significant role in any way to accomplish the work. Grace of the almighty God and blessings of my parents have formed the path to reach my desire goal.

6. REFERENCES

Papers:

Papers:

[1] Neha Goel , Dr. C.K.Jha “Preprocessing Web logs: A Critical phase in Web Usage Mining ”, International Conference on Advances in Computer Engineering and Applications (ICACEA) IMS Engineering College, Ghaziabad, India, ISSN: 1530 9866,pp. 672 – 676,2015.

[2] Avadh Kishor Singh, Ajeet Kumar, Ashish K. Maurya ,“Association Rule Mining for Web Usage Data to Improve Websites”, International Conference on Advances in Engineering & Technology Research (ICAETR), pp- 1-6,ISSN : 2347-9337,2014

[3] Avadh Kishor Singh, Ajeet Kumar, Ashish K. Maurya“An Empirical Analysis and Comparison of Apriori and FP- Growth Algorithm for Frequent Pattern Mining ”, International Conference on Advanced Communication Control and Computing Technologies (lCACCCT) ,pp-1599 - 1602,ISSN: 1487 0338,2014

[4] Murli Manohar Sharma , Anju Bala “An Approach for Frequent Access Pattern Identification in Web Usage Mining”, Advances in Computing, Communications and Informatics (ICACCI), 2014 International Conference on 24-27 Sept. 2014,pp- 730 - 735,ISBN- 978-1-4799-3078-4.

[5] Mr.N.P.Jilhedar ,Dr.S.K.Shirgave “User Web Usage Mining for navigation improvisation using semantic related frequent patterns”International Conference On Computer and Communications Technologies (ICCCT), pp- 1-5,ISSN-1502 2122,2014

[6] Akshita Bhandaria, Ashutosh Guptaa, Debasis Dasa,“Improvised apriori algorithm using frequent pattern tree for real time applications in data mining ”, International Conference on Information and Communication Technologies (ICICT) ,Elsevier,2015

[7] Sudhir Tirumalasetty, Sreenivasa Reddy Edara“A New Algorithm in Association Mining, Amoeba for Finding Frequent Patterns Using Functional Dependency and Probability ”, International Conference on Information and Communication Technologies (ICICT),pp-31-36, Elsevier,2015

[8] Liu Kewen “Analysis of Preprocessing Methods for Web Usage Data” 1nternational Conference on Measurement, Information and Control (MIC),ISSN- 978-1-4577-1601-0,pp-383-386,2012.

[9] R. Agrawal, T. Imielinski and A. Swami, “Mining association rules between sets of items in large databases” In Proceeding of the ACM SIGMOD International Conference on Management of Data (ICMD), USA, pp. 207–216, 1993.

[10]Cooley, R., Mobasher, B., and J. Srivastava, “Grouping web page references into transactions for mining World Wide Web browsing patterns” Proceeding of the IEEE Knowledge and Data Engineering Exchange Workshop (KDEEW), Newport Beach, CA, pp 2-9, ISSN:5781 6676 ,1997.

[11]E.R. Omiecinski, “Alternative interest measures for mining associations in databases” IEEE Transactions on Knowledge and Data Engineering, vol.15, Issue 1, pp. 57-69, ISSN : 1041-4347, 2003.

[12]L.Cristofor and D.Simovici, “Generating an informative cover for association rules” Proceeding of the IEEE International Conference on Data Mining (ICDM), Boston, USA, pp. 597- 600,ISSN: 7695-1754, 2002.

[13]A.Saleem Raja, E.George and Dharma Prakash Raj “MAD- ARM: Mobile Agent based Distributed Association Rule Mining” IEEE International Conference on Computer Communication and Informatics (ICCCI) Coimbatore, pp. 1- 5, ISSN: 1335-7197,2013.

Websites:

[14] https://en.wikipedia.org/wiki/Precision_and_recall at 6:29 am Friday December 11, 2015.

[15]https://en.wikibooks.org/wiki/Data_Mining_Algorithms_In_R/Frequent_Pattern_Mining/The_FP-Growth_Algorithm at 10:30pm Monday February 29,2016

Book:

[16] Jiawei Han,Micheline Kamber,Jian Pei, Data Mining Concepts and Techniques,Morgan Kaufmann Publishers