Improved Technique to Discover Frequent Pattern Using FP-Growth and Decision Tree

[1]Meera J. Tank, [2]Firoz A. Sherashiya
[1]PG Student, [2]Assistant Professor
[1] Department of computer engineering,
[1]Darshan institute of Engineering and Technology, Rajkot,Gujarat, India.

***Abstract*— Web usage mining is the application of web mining, which implements various techniques of data mining for discovery and analysis of patterns in click stream and associated data collected or generated as a result of user interactions with web resources on one or more web sites. It consists of three phases which are data Preprocessing, pattern discovery and pattern analysis. In the pattern analysis phase interesting knowledge is extracted from frequent patterns and these results are used for website modification. In the proposed paper, a hybrid approach is used to fetch HTML as well as XML contents from a web page. In this approach combined effort of FP-Growth algorithm and Decision Tree is applied for pattern discovery. This approach helps in finding effective usage patterns. FP-Growth algorithm is used to remove the unimportant information from the contents and Decision tree is used to fetch the contents from a web page.**

***Index Terms* - Web Mining, Apriori, FP-Tree, Decision Tree**

I. INTRODUCTION

WWW is the largest source of information and Information is growing at a rapid rate. With the growth of explosive Internet information, the data mined from the web are useful information, which has gradually become a very important research as the number of documents grows, searching for information is turning into a cumbersome and time consuming operation. Web Mining plays a pivotal role in extracting patterns of related and unrelated information and knowledge. Web Mining is the application of data mining which is used to generate patterns from the web. Patterns must be such that they are easily understandable, useful and novel. Not only data mining but also other tools from fields of artificial intelligence, machine learning, natural language processing can also be used efficiently to fetch web data. Web Mining on the basis of type of data to be explored can be divided into three main categories.

A. Web usage mining

Web usage mining also known as web log mining, aims to discover interesting and frequent user access patterns from web browsing data that are stored in web server logs, proxy server logs or browser logs. By web usage mining, commercial websites take advantage of knowing the usage pattern of customer, their behavior and frequency of their visits [2].

B. Web structure mining

Web structure mining operates on the Web's hyperlink structure.Web structure mining is the process of using graph theory to analyze the node and connection structure of a web site. It focuses on the connectivity of the web site to other sites that are called as hyperlinks [2].

C. Web content mining

Web content mining is the process to discover useful information from text, image, audio or Video data semi-structured records in the form of html and xml contents which are either embedded in the web page Or having links to other pages in the web. Web content mining sometimes is called web text mining, because the text content is the most widely researched area.

Since HTML has many limitations like limited tags, not case sensitive and designed to display data only, Web developers has started to develop Web pages on emerging Web Technologies like XML, Flash etc. XML was designed to describe data and to focus on what the data is. XML also plays the role of a met language and allows document authors to create customized mark-up language for limitless different types of documents, making it a standard data format for online data exchange. This growing use has raised need for better tools and techniques to perform mining on XML too. In the proposed paper, a combine approach is used to fetch XML contents from XML file. Rest of the paper is organized as follows [1].

II. LITERATURE REVIEW

In [1], **Rupinder Kaur[1], Kamaljit Kaur[2]** provides a hybrid approach used to removing the unimportant information from the contents and fetch the contents from a web page. Hyperlinks and images are fetched from the web using two well know data mining algorithms named Apriori and Decision Tree. These algorithms when applied individually previously gave more reliable results. So, that's why these algorithms are chosen and a combined approach is proposed.

In [3], **Latika Tamrakar[1], S. M. Ghosh[2]** used Apriori algorithm for the discovery of most frequent associated pages. so that most frequent navigation pages can be retrieved. Pattern analyzer can use these patterns for performing some important applications like system Improvement by page caching, site modification, page personalization, website restructuring etc.

In [4], **Mr. Rahul Mishra, Ms. Abha Choubey** compares Apriori and FP-growth, which shows advantage of FP-growth over Apriori. The main drawback of Apriori algorithm is that the candidate set generation is costly, especially if a large number of patterns and/or long patterns exist. The FP-growth algorithm is one of the fastest approaches for frequent item set mining. The FP-growth algorithm uses the FP-tree data structure to achieve a condensed representation of the database transaction and employees a divide-and conquer approach to decompose the mining problem. Our experimental result shows that the FP-growth method is efficient and scalable for mining both long and short frequent patterns.

In [5], **Bhaiyalal Birla, Sachin Patel,** shows the limitation of Apriori algorithm. ***Drawback of apriori are,***

- Found that when the size of data increases the developed model perform poor results and some are provide more accurate results.
- In Frequent set mining our modify Apriori algorithm takes less time than Apriori algorithm for search patterns and building data models.
- Memory used is directly proportional to the size of data in both kinds of algorithms used in web mining.
- Accuracy is not much depends on the size of data it is most of the time depends upon type of data.

In [6], **Djellel Eddine Difallah, Ryan G. Benton, Vijay Raghavan, Tom Johnsten** compare Apriori and FP-growth, which shows advantage of FP-growth over Apriori. The FP-Growth algorithm is a divide–and–conquer approach that is considered to be an order of magnitude faster than Apriori. It relies on a special tree data structure called FP-Tree which is obtained by ordering the transaction attributes values by their frequency, pruning those that do not meet a given minimum support, and then inserting the transaction, or action sets in our case, into a tree. The result is a condensed data structure that avoids expensive database scans and is especially tailored for dense datasets. The size of the FP-Tree constructed from the action table in the Nursery experiment with varying minimum support values. While the action table size is always 250Mb, the size of the FP-Tree is 3.5Mb even for very small minimum support. The size of the FP-Tree shrinks even further for increasing minimum support values and this is due to the pruning at this level frequent association action rules mining using FP-Tree (FAARM) has a better execution time on relatively small dataset, over ARD and AAR.

In [7], **Bo Wu, Defu Zhang, Qihua Lan, Jiemin Zheng,** shows advantage of FP-growth over Apriori Algorithm. Most algorithms were based on Apriori algorithm which generated and tested candidate itemsets iteratively. This may scan database many times, so the computational cost is high. In order to overcome the disadvantages of Apriori algorithm and efficiently mine association rules without generating candidate itemsets, a frequent pattern- tree (FP-Growth) structure is proposed in.

A great advantage of FP-tree is that overlapping itemsets share the same prefix path. So the information of the data set is greatly compressed. It only needs to scan the data set twice and no candidate itemsets are required.

III. PROPOSED APPROACH

In this approach, two well-known data mining algorithms are used. Our approach is applied in two steps-

Step 1: In the first step Decision tree is implemented to fetch the data from the web.
Step 2: FP-Growth algorithm is used to remove the unimportant information from the contents.

Step 1: *Decision Tree*

A decision tree is used for decision making purpose. Decision tree has root and branch node. From the root node,

Users split each node recursively based on decision tree learning algorithm. The final result of decision tree consists of branches and each branch represents a possible scenario of decision and its consequences [1].

Input

- Data partition, D, which is a set of training tuples and their associated class labels.
- Attribute list, the set of candidate attributes.
- Attribute selection method, a procedure to determine the splitting criterion that best partitions the data tuples into individual classes. This criterion includes a splitting attribute and either a splitting point or splitting subset.

Output:

A Decision Tree

Step 2: *FP-Growth algorithm* [8]

In [8], Han, Pei et al. proposed a data structure called FP-tree (frequent pattern tree). FP-tree is a highly compact representation of all relevant frequency information in the data set. Every path of FP-tree represents a frequent item set and the nodes in the path are stored in decreasing order of the frequency of the corresponding items. A great advantage of FP-tree is that overlapping item sets share the same prefix path. So the information of the data set is greatly compressed. It only needs to scan the data set twice and no candidate item sets are required.

Input:

Decision Tree

Output:

Remove the unimportant information from the contents

IV. METHODOLOGY

It describes the process of fetching web contents like hyperlinks using the hybrid approach. Efficient design architecture of the proposed Hybrid method is shown in figure1.And workflow of architecture is shown in figure2.

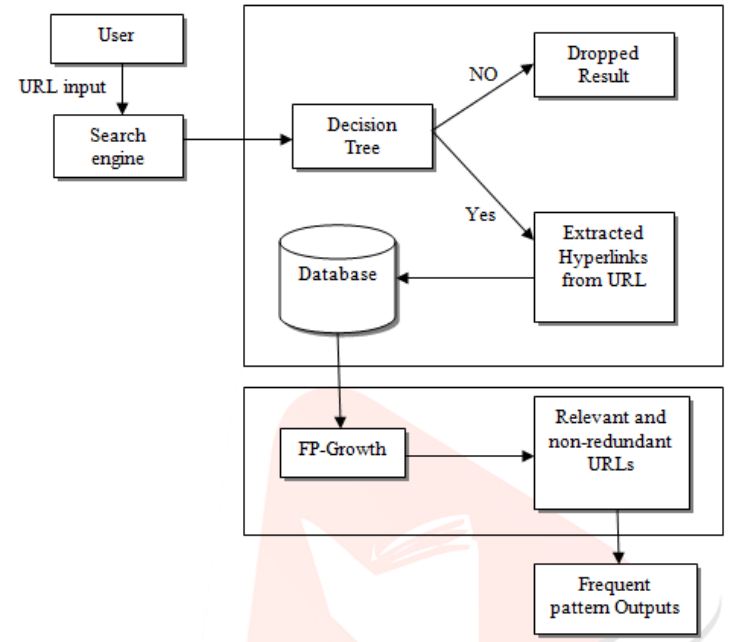

Figure1. Architecture of hybrid approach

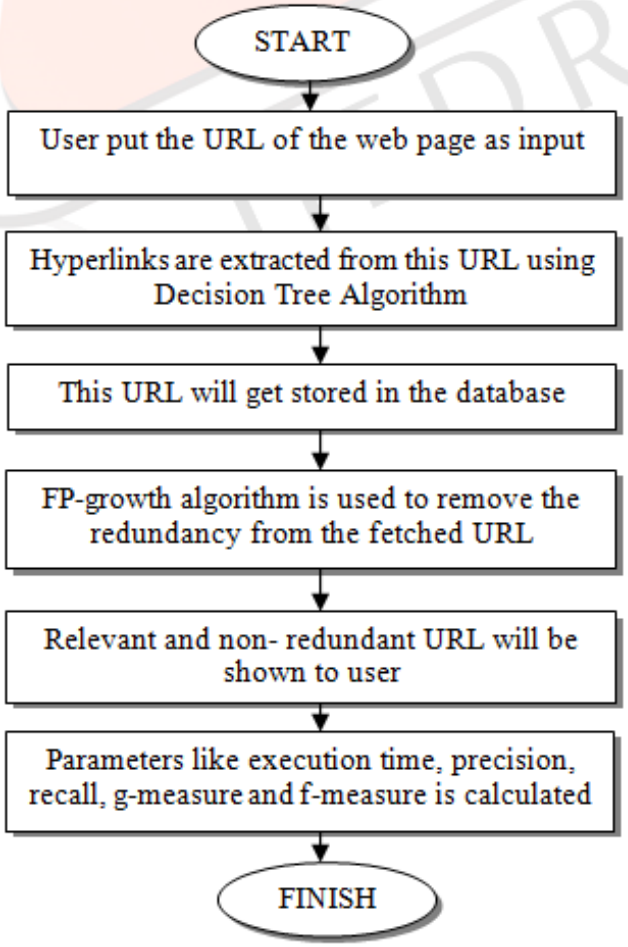

Figure 2.Architectural flowchart

V. CONCLUSION

In this proposed approach, hyperlinks are fetched from the web using two well know data mining algorithms named FP - Growth and Decision Tree. FP-Growth algorithm is used to remove the unimportant information from the contents and Decision tree is used to fetch the contents from a web page. Main advantage of this algorithm is it is more efficient and less time required to fetch the data.

VI. FUTURE SCOPE

Though, our proposed approach implements basic classic algorithms which are FP-growth and Decision Tree Induction algorithm. There are several modifications proposed to these algorithms like Hybrid FP-growth and for Decision Tree certain other algorithms are there like C4.5, CART which can be implemented in future to increase the efficiency of this hybrid approach. Also this approach runs best while using high speed internet, some methods can be proposed in future so that this approach can give its best at slow speed internet too.

REFERENCES

[1] Rupinder Kaur[1], Kamaljit Kaur[2], "An Improved Web Mining Technique to Fetch Web Data Using Apriori and Decision Tree" International Journal of Science and Research (IJSR) Volume 3 Issue 6, June 2014.

[2] Shaily G.Langhnoja[1], Mehul P. Barot[2], Darshak B. Mehta[3] , "Web Usage Mining Using Association Rule Mining on Clustered Data for Pattern Discovery" International Journal of Data Mining Techniques and Applications Vol 02, Issue 01, June 2013.

[3] Latika Tamrakar[1], S. M. Ghosh[2] , "Identification of Frequent Navigation Pattern Using Web Usage Mining" International Journal of Advanced Research in Computer Science & Technology (IJARCST 2014) 296 Vol. 2, Issue 2, Ver. 2 (April - June 2014).

[4] 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

[5] Bhaiyalal Birla, Sachin Patel, "An Implementation on Web Log Mining" International Journal of Advanced Research in Computer Science and Software Engineering Volume 4, Issue 2, February 2014

[6] Djellel Eddine Difallah, Ryan G. Benton, Vijay Raghavan, Tom Johnsten, "FAARM: Frequent association action rules mining using FP-Tree" 2011 11th IEEE International Conference on Data Mining Workshops

[7] Bo Wu, Defu Zhang, Qihua Lan, Jiemin Zheng, "An Efficient Frequent Patterns Mining Algorithm based on Apriori Algorithm and the FP-tree Structure" 2008 3rd IEEE International Conference on Convergence and Hybrid Information Technology

[8] J. Han, and M. Kamber, "Data Mining Concepts and Technique". Morgan Kaufmann Publishers, 2000

[9] http://en.wikipedia.org/wiki/Web_content

[10] http://webdesign.about.com/od/content/qt/what-is-web-content.htm

[11] http://www.tutorialspoint.com/data_mining/dm_dti.htm

[12] http://en.wikipedia.org/wiki/Decision_tree

[13] http://en.wikipedia.org/wiki/Association_rule_learning