An Approach to Enhance Web Service Resource Framework using the Improved PLWAP Algorithm for Large Scale Hybrid Data in Distributed Environment

Ashish Sharma, Niket Bhargava

Abstract—With the accessing of millions of web pages for business and personal transactions, huge amounts of web page access data have been stored in web servers. The existing web pattern mining systems do not provide the necessary tools and functionality to capture all stored distributed web page sequences easily. Data mining represents the process of extracting interesting and previously unknown knowledge from data. Through the approach described in this paper we address the important data mining problem of sequential web pattern access. We present an approach for mining the sequential web patterns with Distributed mining model. Here the environment under consideration is the one where raw input data is collected from number of web servers at distributed locations. This approach is a combination of distributed concept for web personalization. It utilizes the concept of WAP algorithm and provides ease of getting web sequential patterns enhancing “Web Service Resource Framework”.

Index Terms—Distributed data mining, WAP algorithm, Web access Patterns, Web Personalization.

I. INTRODUCTION

Data mining refers to extracting or “mining” knowledge from large amounts of data. Data mining should have been more appropriately named knowledge mining from data. There are many other terms carrying a similar or slightly different meaning to data mining, such as knowledge mining from databases, knowledge extraction, data/pattern analysis, data archaeology, and data dredging [13]. One of the techniques of data mining is Web Mining which is the extraction of interesting and potentially useful patterns and implicit information from artifacts or activity related to the Worldwide Web. There are roughly three knowledge discovery domains that pertain to Web mining as Web Content Mining, Web Structure Mining, and Web Usage Mining. Web content mining is the process of extracting knowledge from the content of documents or their descriptions. Web document text mining, resource discovery based on concepts indexing or agent based technology may also fall in this category. Web structure mining is the process of inferring knowledge from the Worldwide Web organization and links between references and referents in the Web. Finally, web usage mining, also known as Web Log Mining, is the process of extracting interesting patterns in web access logs [26]. The Web usage mining is a process of extracting useful information from server logs i.e. user’s history. Web usage mining is the process of finding out what users are looking for on the Internet. Some users might be looking at only textual data, whereas some others might be interested in multimedia data [31] [32]. The problem of discovering sequential patterns [Mannila 1995, Srikant 1996] [5] in the domain of web mining is to find inter-transaction patterns such that the presence of a set of items is followed by another item in the time-stamp ordered transaction set. In Web server logs, a visit by a client is recorded over a period of time. The time stamp associated with a transaction in this case will be a time interval which is determined and attached to the transaction during the data preprocesses.

The discovery of sequential patterns in Web server access logs allows Web-based organizations to predict user navigation patterns and helps in targeting advertising aimed at groups of users based on these patterns. Sequential Pattern Mining comes in Association rule mining. For a given transaction database $T$, an association rule is an expression of the form $X \cup Y$, where $X$ and $Y$ are subsets of $A$ and $X \cup Y$ holds with confidence . if % of transactions in $D$ that support $X$ also $Y$. The rule $X \cup Y$ has support in the transaction set $T$ if % of transactions in $T$ supports $X \cup Y$. Association rule mining can be divided into two steps. Firstly, frequent patterns with respect to support threshold min sup are mined. Secondly association rules are generated. The remaining paper is organized as follows: Section 2 explained a brief review of the researches related to the proposed system. Section 3 presents the proposed Data Mining Model in Distributed environments. The experimental results of the proposed system are given in Section 4 and the conclusions are coming up in Section 5.

II. LITERATURE REVIEW

Association Rule Mining [4] as the technique for using concepts of Web Usage Mining. Association Rules are probably the most elementary data mining technique and, at the same time, the most used technique in Web Usage Mining.

90
When applied to Web Usage Mining, association rules are used to find associations among web pages that frequently appear together in users’ sessions. The typical result has the form “A.html, B.html -> C.html” which states that if a user has visited page A.html and page B.html, it is very likely that in the same session, the same user has also visited page C.html. Learning web usage patterns proposes and evaluates some interestingness measures to evaluate the association rules mined from web usage data. Now Association Rule Mining can be implemented using 3 Techniques: FP-Tree, WAP-Tree, and PLWAP-Tree.

FP-Tree (Frequent Pattern Tree) is a variation of tree data structure, which is a Postfix-tree structure for storing Crucial & Compressed Information. Thus it is possible to alleviate Multi-Scan Problem using Above Information. A WAP-tree is an aggregate tree that represents the web access sequence database. All nodes with the same label are linked when the tree is built. In PLWAP-Tree, it was proposed to link all the nodes of same label in the order determined by a pre-order traversal of the tree after it is built. Such a tree is, thus, called Pre-Order Linked WAP-tree (PLWAP-tree). The mining algorithms in both WAP-tree and PLWAP-Tree use recursive conditional searching of projection databases to find frequent web access patterns. The difference is that the WAP-tree mining grows the suffix of frequent patterns, while the PLWAP-tree mining grows the prefix of the frequent patterns. It is shown in that the mining time of the PLWAP-tree algorithm is less than that of the WAP-tree algorithm and the GSP algorithm. There are three modules in the project, they are Filtering/Cleaning of Raw web logs and Compatibility Mode.

Implementation of PL-WAP Algorithm

Unlike the conditional search in WAP-tree mining, which is based on finding common suffix sequence first, the PLWAP technique finds the common prefix sequences first. The main idea is to find a frequent pattern by progressively finding its common frequent subsequences starting with the first frequent event in a frequent pattern. For example, if abcd is a frequent pattern to be discovered, the WAP algorithm progressively finds suffix sequences d, cd, bcd and abcd. The PLWAP tree, on the other hand, would find the prefix event a first, then, using the suffix trees of node a, it will find the next prefix subsequence ab and continuing with the suffix tree of b, it will find the next prefix subsequence abc and finally, abcd. Thus, the idea of PLWAP is to use the suffix trees of the last frequent event in an m-prefix sequence to recursively extend the subsequence to m+1 sequence by adding a frequent event that occurred in the suffix trees. Using the position codes of the nodes, the PLWAP is able to know the descendant and sibling nodes of oldest parent nodes on the suffix root set of a frequent header element being checked for appending to a prefix subsequence if it is frequent in the suffix root set under consideration.

III. PROPOSED WORK

Large databases, sometimes distributed over several remote locations, are becoming more common in the contemporary Global Economy scenario. The local databases which were initially small, have grown, growing continually and getting distributed to several remote sites as a result of globalization. Many of the conventional data mining algorithms are ineffective and inefficient for handling large and distributed data sets. Hence, the scalability of data mining algorithms has become an active area of research with many challenging problems. Here we are proposing a method for enhancing the web service resource framework by enabling the conventional sequential pattern accessing methods to distributed environments having large scale data sets of hybrid nature situated different distributed locations in the semantics web. This approach uses client-server based distributed data mining concept in which at any central node we have to collect the mining information from various distributed nodes for generating the integrated output of sequential accessing of patterns that can be used in various
area of applications. We also proposing a new algorithm which would be able to handle large amount of industrial (hybrid i.e. data collected from various web server located in distributed fashion for mining input) data set for scaling this approach to a broader way. The proposed approach is for accessing the sequential web access patterns from data set received from various web server logs running in distributed manner having hybrid or heterogeneous web log information. The proposed method uses the PLWAP algorithm for mining the sequences from input data. The following changes have been made in original PLWAP.

- The algorithm is changed to handle large scale of data sets rather than taking the lab simulated sample input data set,
- The algorithm is modified to handle different types of web logs i.e. the web logs treated as the input data set are not exactly same for each iteration.

Another changed in the approach is proposed to enhance the flexibility, scalability, availability and transparency that is the data mining process is distributed rather than centralized, means in this distributed data mining process for collecting the sequential web access patterns for the purpose web personalization over large scale of hybrid data the Modified PLWAP algorithm is running on number of independent nodes generating the sequential web access patterns locally. After getting the local web access sequence at each independent node a file collection and result integration module that is suppose to remove the ambiguous result entries and to sorting the final result is used. In this section we discussed about some of the basic techniques of distributed processing of large datasets, for data mining tasks and adaptability of some core data mining algorithms (such as PLWAP) for distributed processing. Contemporary research work related to distributed processing of the data mining algorithms follows next.

There are following stages in developing the D-PLWAP system [PROPOSED ALOGORITM].

//BEGIN
[1] INPUT=Large scale heterogeneous web log data set as a raw input at each node [LWASD].
[2] Call Filtering/cleaning module at each individual node.
[3] Converting the filtered large data set in a compatible mode at each individual node.
[4] Call PLWAP for large data set at each node individually (D-PLWAP) with MS value 0-1.
[5] Call DPLWAP module at any selected node (Decision node).
// END
The following figure3 is giving an overview of entire component.

![Fig 3 Module for Final Result generation (Decision Node)](image)

IV. RESULT ANALYSIS

Here in this section we will explain the procedure adopted to experiment the proposed approach with respect to a older one named ODAM (Optimal Distributed Association Rule Mining). Here the input data set used for experimental purpose is considered of different sizes like 5000, 10000, 15000, 20000 records in all the files to be processed. The component/module responsible to generate the Global sequential access patterns have to process the above sized files as well as the ODAM would also process the same for comparative study. The results for both the approaches for processing the files of various sizes at a node running File Processing module are given below for toe different environments showing the speed up of the data mining task on increasing the number of distributed nodes for large numbers of records. The results for both the scenarios are shown below.

**Result for Single PC**

<table>
<thead>
<tr>
<th>TABLE 1 Comparison of results for Single PC Environment</th>
</tr>
</thead>
<tbody>
<tr>
<td>RECORD</td>
</tr>
<tr>
<td>ODAM</td>
</tr>
<tr>
<td>D-PLWAP</td>
</tr>
</tbody>
</table>
The above sized files having different number of records are given to the system under three considerations i.e. on single computer system, on two computer system and on five number of computer systems. Here, we have supposed equal distribution of data in all machines. For every set of data with different record numbers, depending on the number of participating machines in the mentioned procedure, we will come to different results. If we examine the obtained results, we will see that by doubling the number of calculating resources for large data sets in Grid, the execution time will decrease to lower than that of the previous one. We can see that in smaller sets of data; the increase in the number of calculating resources will not improve the efficiency of the procedures. But in some cases will reduce the efficiency. The reason is the delay which occurs in transmission of data and its processing in coordinates.

V. CONCLUSION AND FUTURE WORK

Our first contribution represents the general methodology for preprocessing described in earlier. This methodology consists in four data preprocessing steps: data fusion, data cleaning, data structuring and data summarization. The first three steps are the classical preprocessing steps while the last
one represents an advanced preprocessing step we created for summarizing and aggregating the structured data. Large log files coming from Web servers are used as an input to preprocessing component at each node.

Then, unnecessary requests are removed from this unique log file during the data cleaning step. The remaining requests are grouped in user sessions, visits and episodes and saved to either structured flat log file or a relational database designed according to our model. In the last step of data preprocessing, the data summarization, summary variables are computed at different detail levels such as user session or visit. These variables are stored in the database and will be used in the data mining step that follows. The work is still going on for generalization. This work also adapts the PLWAP-tree structure in different manner for storing frequent sequential patterns to be mined. However, to enrich the utilities and to propose a system for distributed data mining we perform the execution of PLWAP at number of distributed nodes (called Distributed –PLWAP) generating the local Sequential web access patterns. These number of local sequential web access patterns are to given a file processing system (called DPLWAP file processing system) written in java language for generating the global sequential web access patterns. The proposed system is compared with Optimal Distributed Association Rule Mining (ODAM) system for validating and verifying the proposed system. For mining sequential patterns from web logs, the following aspects may be considered for future work. The Proposed system could be extended to handle sequential pattern mining in large traditional databases other than web log and any other order can be consider for improvement pre-order linkage. Efficient web usage mining could benefit from relating usage to the content of web pages. Other areas of interest for future work include proposed work with bioinformatics and Artificial Intelligence. Also this work can be couples with parallel approach for enriching the architectures for Grid computing, Cluster computing and cloud computing. We can apply these techniques to incremental mining of web logs and sequential patterns.

The future work also includes the proposal to reduce the communicational overheads and latencies. Also making the provisions for more interactive and fast file processing system for the speed up of the entire process. The extraction of Sequential web access patterns may also be customize for the purpose of mining the information utility wise and to predicting the nature of next sequence called Advanced Web Personalization.

REFERENCES


[17] C.I. Ezeife School of Computer Science, University of Windsor, and Windsor, Ontario, Canada N9B 3P4 ezeife@uwindsor.caMin Chen School of Computer Science, University of Windsor, and Mining Web Sequential Patterns Incrementally with Revised PLWAP Tree.


[19] Zhou Baoyao Intelligent Web Usage Mining, Division of Information Systems School of Computer Engineering.

[20] Ning Chen1, Nuno C. Marques1, and Narasimha Bolloju2, A Web Service-based approach for data mining in distributed environments, CENTRIA, Department of Information FCT, New University of Lisbon, Portugal.

[21] Asgarali Bouyer Islamic Azad University-Malekan Branch, Iran Bahman Arasteh, Mohammad Bager Karimi, Mohammad Javad Hoseyni.Ali Bouyer Islamic Azad University-Sama Branch, Iran Ali Movagher Sharif University Of Technology-Tehran, Iran A New And
Optimal Architecture For Applying Association Rule Mining Algorithms On Grid.

[22] Christie I. Ezeife 1, Monwar Mostafa 1 School of Computer Science, University of Windsor, Windsor, Canada, N9b 3p4 A Wap-Based Algorithm Mining Frequent Sequential Stream Patterns.


[24] C.I. EZEIFE cezeife@uwindsor.ca YI LU School of Computer Science, University of Windsor, Windsor, Ontario, Canada, Data Mining and Knowledge Discovery, 10, 5–38, 2005, Mining Web Log Sequential Patterns with Position Coded Pre-Order Linked WAP-Tree.

[25] Thomas Legler SAP AG69190 Walldorf, Germany, Robust and Distributed TopN Frequent Pattern Mining.


[27] C.I. Ezeife School of Computer Science, University of Windsor, Windsor, Ontario, Canada, Mining Web Sequential Patterns Incrementally with Revised P1WAP Tree 2006.