ADVANCE APPROACH FOR DATA CLASSIFICATION USING DISTRIBUTED ASSOCIATION RULES

Sonali Sonkusare[*]

Mr. Jayesh Surana[**]

Abstract

In the internet era web sites on the internet are useful source of information for almost every activity. So there is a rapid development of World Wide Web in its volume of traffic and the size and complexity of web sites. Web mining is the application of data mining, artificial intelligence, chart technology and so on to the web data and traces user's visiting behaviors and extracts their interests using patterns [Mahendra and Shefalika,2012]. Because of its direct application in e-commerce, Web analytics, e-learning, information retrieval, web mining has become one of the important areas in computer and information science. There are several techniques like web usage mining exists. But all processes its own disadvantages. Pattern mining is a heavily researched area in the field of data mining frequent with wide range of applications. One of them is to use frequent pattern discovery methods in Web log data. Discovering information from Web log data is called Web usage mining. The aim of discovering frequent patterns is to obtain information about the navigational behavior of the user in Web log data.

Keywords- *Periodic, Pattern; Weight constraint; Web usage mining;*

[*] M.E. Student

[**] Assistant Professor

I. INTRODUCTION

It is not overstated to say the internet is the majority excited impacts to the human society in the last 17 years. To predict the users' behaviors and personalize information to decrease the traffic load, the web service providers wants to discover the way. Web Usage Mining is the automatic discovery of user access patterns from Web servers. Organizations collect large volumes of data in operations occurring daily, generated automatically by Web servers that are collected in Web access log files. Web usage mining involves the automatic detection of user access patterns on one or more web servers. It is an application of data mining algorithms to web access logs to find the trends and regularities in web users' navigation patterns. There are many kinds of data that can be used in web mining and can be classified into following five types:

- Content of Web Page.
- Inter-Page Structure of Web Page.
- Intra-Page Structure of Web Page
- Usage Data.
- User Profile.

Web mining can be divided into three research fields, i.e., web content mining, web structure mining, and web usage mining. Web usage mining is the application of established data mining techniques to analyze Web site usage. That data will be useful for a user to predict which Web pages will be clicked based on previous behavior An Efficient Web Mining Algorithm for Web log Analysis: E-Web Miner[Mahendra, Shefalika shefalika,2012] they was proposed only a small data set pattern discovering not working large data set and distributed environment. They focused on E-Web Miner may take into consideration the support and the confidence of any sequential pattern of web pages of users. This may provide further refinement in the result of candidate sets. Graphical presentation of the output is not convincing enough to prove that the Web Miner position much above the other web mining algorithm - a sample of which is represented by Apriori All algorithm but we focus on distributed association rules and more improvement Apriori algorithm. In this research we focus on distributed association rules mining method, we propose a An proficient Web Mining Algorithm for Web Log Analysis using Distributed association rules Mining for correct web page prediction. Some of the existing systems used for web traversal based on his or her navigational behavior on the web. The proposed system will use visiting frequency of a page, time spent on a page to assign a quantitative weight to each page for a user. The instinct of this approach is that the time spent on pages and visiting frequency is biasing factors to illustrate the interest on a page. The other systems give user recommendations related to a navigation session or the user profile stored in the system. Section should provide the concept and overview of the domain of proposed Work giving the references of work being in the field under same context.

II. RELATED WORK

The problem of discovering association rules was first introduced in [6] and an algorithm called AIS was proposed for mining association rules. For last fifteen years many algorithms for rule mining have been proposed. Most of them follow the representative approach by Agrawal et al. [7], namely Apriori algorithm. Various researches were done to improve the performance and scalability of Apriori included using parallel computing. There were also studies to improve the speed of finding large itemsets with hash table, map, and tree data structures. Here we review some of the related work that forms a basis for technique

Mahendra Pratap Yadav in at al[1]E-Web Miner may take into consideration the support and the confidence of any sequential pattern of web pages of users. This may provide further refinement in the result of candidate set pruning. The graphical presentation of the output is convincing enough to prove that the Web Miner stands much above the other web mining algorithm - a sample of which is represented by Improved AprioriAll algorithm.

S.Veeramalai in at al[2]With modified Apriori algorithm the structure is formulated with the help of hash tree algorithm. Design tool allows experimenting with the concepts of fuzzy modify association rules. It finally analyzed the crisp boundary problem in the combined algorithm and it is overcome by modified association Apriori hash tree fuzzy algorithm and the efficiency is increased in it.

Jiyi Xiao in at al[4] proved that the GAHMM training gave better quality of solutions than the Baum-Welch algorithm by optimizing the HMM model parameters for the HMM training. In this paper, author describe GA-HMM training in that they finds the optimal number of states for the web model and also optimizes the model parameters in a single step. In addition, they combine the GA with the Baum-Welch algorithm to form a hybrid-GA such that the quality of our results and the runtime behavior of the GA are improved.

Sang T.T. Nguyen in at al[5]proposed a new mining algorithm for WUM based on the Markov model from the stochastic approach. The new feature of this algorithm is that it predicts users' web navigation patterns using the frequent web access sequences extracted from the web logs rather than all web pages. As a result, the complexity of the Markov model can significantly decrease because only frequently accessed web pages are used, resulting in a small number of states. The frequent web access sequences fed into the model can be discovered from web log data by using a tree algorithm.

AIS Algorithm: The main drawback of the AIS[5] algorithm is that it makes multiple passes over the database. Furthermore, it generates and counts too many candidate itemsets that turn out to be small, which requires more space and wastes much effort that turned out to be useless.

Apriori Algorithms: Generally, an association rules mining algorithm contains the following steps [8]:

The set of candidate k itemsets is generated by 1-extensions of the large (k 1) itemsets generated in the previous iteration. Supports for the candidate k itemsts are generated by a pass over the database. Itemsets that do not have the minimum support are discarded and the remaining itemsets are called large k itemsets. This process is repeated until no more large itemsets are found

Partition Algorithm: The Partition algorithm [9] logically partitions the database D into n partitions, and requires just two database scans to mine large itemsets.

DHCP Algorithm: The DHP (Direct Hashing and Pruning) algorithm is an effective hash based algorithm for the candidate set generation. It reduced the size of candidate set by filtering any k itemset out of the hash table if the hash entry does not have minimum support. They improved the performance of the conventional Apriori algorithm that mines association rules by presenting fast and scalable algorithm for association rules discovery in large databases and propose a proficient Web Mining Algorithm for Web Log Analysis using Distributed association rules Mining propose an effective Web log mining system that deals with log preprocessing, sequential pattern mining, and result visualizing. Implement a visualization tool to interpret mining results and predict users' future behavior.

III. PROPOSED APPROACH

We Apply the propose algorithm to more extensive empirical evaluation. Our develop approach use real data like retail sales transaction and medical transactions to confirm the experimental results in the real life domain. We Mines the distributed association rules from relational databases and data warehouses (these rules involve more than one dimension or predicate, e.g. rules relating what a customer shopper buy as well as shopper's occupation).

For dealing with the problem of discovering hidden information from large amount of web log data collected by web servers, we Mine the distributed association rules from transaction databases (these rules involve items at different levels of abstraction). Our contribution introduces the process of web log mining, and we illustrate how frequent pattern discovery tasks can apply on the web log data in order to obtain useful information about the user's navigation behavior.

In this scheme process work in parallel on a shared- nothing architecture. The processors begin with counting local support for each item. This counts are then exchanged across the group in order to each processor calculates their sum. After discarding globally infrequent items, each processor constructs the *F-list* structure sorted by descending order of frequent item supports figure 1. In the next phase, local *CFP-trees* are built by scanning local data partitions and considering only local items belonging to *F-list*.

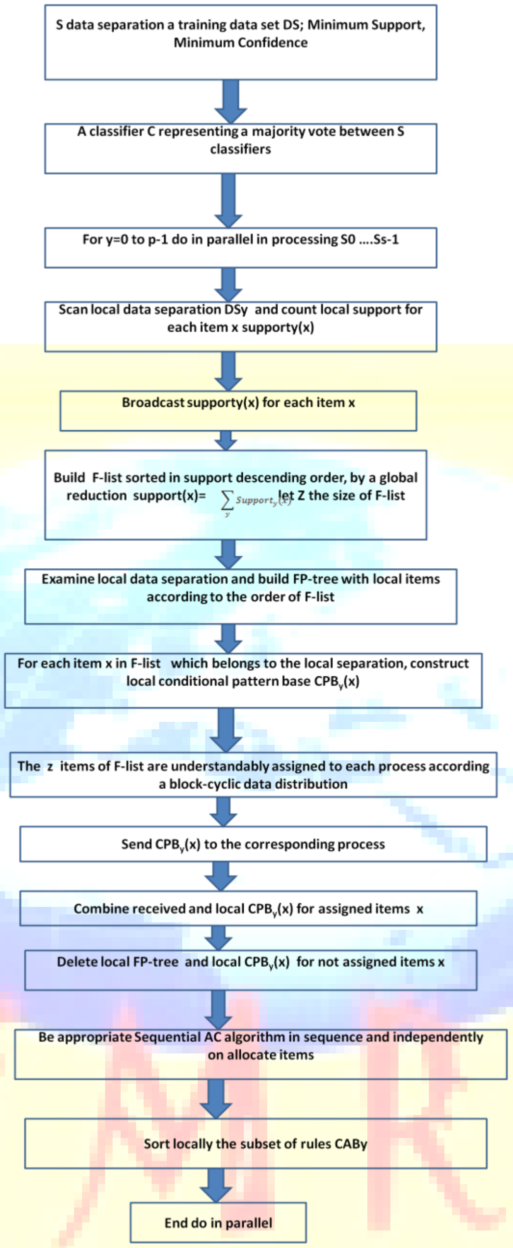
S data separation a training data set DS; Minimum Support, Minimum Confidence
A classifier C representing a majority vote between S classifiers
For y=0 to p-1 do in parallel in processing S0 ….Ss-1
Scan local data separation DSy and count local support for each item x supporty(x)
Broadcast supporty(x) for each item x
Build F-list sorted in support descending order, by a global reduction support(x)= $\sum_y Support_y(x)$ let Z the size of F-list
Examine local data separation and build FP-tree with local items according to the order of F-list
For each item x in F-list which belongs to the local separation, construct local conditional pattern base $CPB_y(x)$
The z items of F-list are understandably assigned to each process according a block-cyclic data distribution
Send $CPB_y(x)$ to the corresponding process
Combine received and local $CPB_y(x)$ for assigned items x
Delete local FP-tree and local $CPB_y(x)$ for not assigned items x
Be appropriate Sequential AC algorithm in sequence and independently on allocate items
Sort locally the subset of rules CABy
End do in parallel

Figure 1: Algorithm DAR

The class label is attached to the last node in the path.Thereafter, the local *CFP-trees* are used in parallel to generate local conditional pattern bases *CPBs* in each processor, for each item in *F-list* figure 1. This partial information will be communicated between the process and merged to generate the initial global *CPB* for each item. A trivial method to assign tasks to different association rules is to perform a bloc-cyclic item distribution, so that the item *x* will be assigned to the numbe. Thereby, each item set will send its local *CPBs* to corresponding process avoiding to send “all” to “all” .

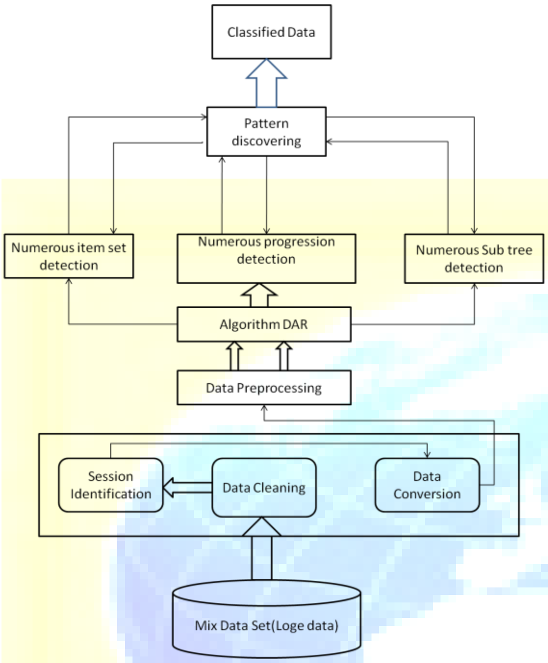

Figure 2: Log classification process

This data distribution strategy could “contribute” to balance the load between the rules, because generally the amount of work needed to process an item of F-list increases for the items with low supports. After this communication,each machine independently mines recursively its assigned items, without any need of synchronization. At this level, several data structures can be deleted from distributed main memories: local *CFP-trees* and local *CPBs* for the items assigned to other task. At the end, if each site products a subset of CARs, the union of all subsets must be exactly the total rule set obtained in the serial version of the algorithm. To classification new record DS, a simple and intuitive technique consists in performing a majority vote between the S components of the final composite model. This strategy was used successfully in ensemble learning methods, with e.g. decision trees as base classifier [10].

The instance to classify is presented to each classifier Cy, (y= 1..s), like in the sequential version, thus it will be classified according to the prediction of the majority, denoted by C = argmax where c is a class label, C_y is the classifier obtained in the processor y, and I(A) is an indicator function that returns 1 if A is true and 0 otherwise. The example in (fig.1) illustrates the parallel construction of F-list. After communicating local counts, itenset compute the sum and discard non global frequent items. Instead of building a global FP-tree, which may not fit in main memory, local FP-trees are constructed in each processor, with the same method.

The support and confidence are the most popular measures for sequential patterns. The support evaluates frequencies of the patterns and the confidence evaluates frequencies of patterns in the case that sub-patterns are given. These parameters are meaningful and important for some applications. However, in other applications, the number of occurrences (support) may not always represent the significance of a pattern. Sometimes, a large number of occurrences of an expected frequent pattern may not be as interesting as few occurrences of an expected rare pattern. This pattern called surprising pattern instead of frequent pattern. The information gain metric which is widely used in the information theory field, may be useful to evaluate the degree of surprise of the pattern. Target is finding set of patterns that have information gain higher than minimum information gain threshold. that the support threshold has to be set very low to discover a small number of patterns with high information gain. Note that surprising patterns are anti-monotonic. It means advantage of standard pruning techniques such as Apriori property can not be used. For example, the pattern ($p_{1,}$ p_2) may have enough information gain while neither (p1,*) nor (*,p_2) does. Given a pattern p= ($p_{1,}$ $p_{2\ldots}$ p_n) and an information gain threshold min-gain, the goal is to discover all patterns whose information gain in the sequence S exceed the min-gain value. Similar to other parameters in data mining algorithms, the appropriate value of the min-gain is application dependent and may be defined by a domain expert. There are some heuristics and methods that user can set the value of this threshold.

Information gain of pattern P is defined as follows:

pg(P)=p(P)*Support(P)

$I(P)=I(p_1)+I(p_2)\ldots.I(p_n)$

$I(p_k)=-\log|p|^{prob(pk)}$ is probability that symbol occurs and is number of events in S.

IV. CONCLUSION

The main advantage of the distributed association rules algorithm is that it discovers the small frequent itemsets in a very quick way, thus the task of discovering the longer ones is enhanced as well.

We finding of the user's browsing patterns can help organizations to provide personalized recommendations of web pages according to the current interests of the user. Reduce information overload by suggesting pages that meet the user's requirement. Many distributed algorithms have been proposed for classification algorithms like decision trees, but so far, there are few works in associative classification. In this paper, we have presented a distributed model which allows the class association rules discovery. Our solution embraces one of the fastest known sequential algorithms (FP-growth), and extends it to generate classification rules in a parallel setting. However, since the data set is distributed. global decision making becomes a difficult task. To avoid the replication of data in the sites, we have chosen to communicate the needed information. This exchange is made only in the first level of recursion, allowing each machine to subsequently process all its assigned tasks independently. At the end, a global classifier is built by all discovered rules, and applying a majority vote strategy. In order to evaluate this choices, it is imperative to carry out an experimental evaluation which permits us in the future to analyze several important costs: accuracy, scalability, speedup, memory usage, communication, synchronization, and also the load balancing.

Reference

[1]Mahendra Pratap Yadav, Pankaj Kumar Keserwani, Shefalika Ghosh Samaddar," An Efficient Web Mining Algorithm for Web Log Analysis: E-Web Miner" 1st Int'l Conf. on Recent Advances in Information Technology | RAIT-2012 |.

[2] Ajith Abraham and Xiaozhe Wang," i-Miner: A Web Usage Mining Framework Using Neuro-Genetic-Fuzzy Approach"

[3] S.Veeramalai, N.Jaisankar and A.Kannan ," Efficient Web Log Mining Using Enhanced Apriori Algorithm with Hash Tree and Fuzzy" International journal of computer science & information Technology (IJCSIT) Vol.2, No.4, August 2010.

[4] Jiyi Xiao Lamei Zou Chuanqi Li," Optimization of Hidden Markov Model by a Genetic Algorithm for Web Information Extraction" ISKE-2007 Proceedings.

[5] Sang T.T. Nguyen," Efficient Web Usage Mining Process for Sequential Patterns" iiWAS2009, December 14–16, 2009, Kuala Lumpur, Malaysia

[6]R. Agrawal, T. Imielinski, and A. Swami. "Mining association rules between sets of items in large databases". In Proceedings of the ACM SIGMOD International Conference on Management of Data (ACM SIGMOD '93), pages 207216, Washington, USA, May 1993.

[7] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules in Large Databases," Prof. 20th Int'l Conf. Very Large Data Bases, pp. 478499, 1994.

[8] K. Sotiris, and D. Kanellopoulos, "Association Rules Mining: A Recent Overview. GESTS International Transactions on Computer Science and Engineering", Vol.32 (1), 2006, pp. 7182.

[9] A. Savasere, E. Omiecinski, and S. Navathe. "An Efficient Algorithm for Mining Association Rules in Large Databases". Proceedings of 21th International Conference on Very Large Data Bases (VLDB'95),

[10] N. Chawla, S. Eschrich, L.O. Hall, "Creating Ensembles of Classifiers," IEEE International Conference on Data Mining, pp. 580-581, 2001.