PALM: Preprocessed Apriori For Logical Matching Using Map Reduce Algorithm

Narayan Gowraj
Department of Information Technology
MIT
Chennai, India
narayan_gowraj@rocketmail.com

Srinivas Avireddy
Department of Information Technology
MIT
Chennai, India
tholi1033@yahoo.com

Sruthi Prabhu
Department of Computer Science
MIT
Chennai, India
sruthe92@yahoo.com

***Abstract*--With the recent explosive growth of the amount of data content and information on the Internet, it has become increasingly difficult for users to find and maximize the utilization of the information found in the internet. Traditional web search engines often return hundreds or thousands of results for a particular search, which is time consuming .In order to overcome these problems, we have described the implementation and design of the PALM Algorithm (PREPROCESSED APRIORI FOR LOGICAL MATCHING) in mining information data from the World Wide Web. The PALM-ALGORITHM provides us with a very efficient and simple way for finding related patterns while maneuvering through the internet. The PALM-ALGORITHM is implemented in two steps. The first step includes a Map-Reducing Algorithm which is used to traverse and analyze all the items of a large database and preprocess them using a variable called MINIMUM THRESHOLD SUPPORT to find the INITIAL CANDIDATE SET(C) AND LARGEITEM SET (L). The second step includes a pre-processing algorithm to find both the CANDIDATE(C) and LARGEITEM SET (L) for the further scans.**

Keywords-PALM (Preprocessed Apriori For Logical Matching) Algorithm, Apriori Algorithm, Map Reducing Algorithm and Pattern Matching.

I. INTRODUCTION

The World Wide Web can be treated as the largest database and mining it to get the information which is high in quality and relevant to our information needs is very difficult. Web mining can be broadly defined as the discovery and analysis of useful information from the World Wide Web [1]. Web mining can be divided into three different types, which are Web content mining, Web structure mining, Web usage mining. Web mining includes techniques to search, collate and analyze patterns in the data content of web sites by using traditional data mining techniques and attributes such as clustering, classification, association, and examination of sequential patterns [2]. We devote the main and the important part of this paper in describing the crux of the PALM-Algorithm and we describe the working and the implementation of the same along with the experimental results.

II. RELATED WORK

The concept of Classification, association and clustering are used in the PALM-ALGORITHM and hence we describe about it before we discuss the PALM-ALGORITHM.

2.1 Classification

Classification is a data mining (machine learning) technique used to predict group membership for data instances. Several major kinds of classification method including decision tree induction, Bayesian networks, k-nearest neighbor classifier, case-based reasoning, genetic algorithm and fuzzy logic techniques. Classification maps data into predefined groups or classes. It is often called as supervised learning because the classes are determined before examining the data [3]. The most common Classification problem is the Decision tree induction. A decision tree is a method that consists of a flow-chart like tree structure where each node represents test on an attribute value and each edge represents the result of the test done on that attribute. For example consider a scenario where you want to segregate the system which is most vulnerable to crashes due to the Mellisa and MyDoom virus. From the data given in the Table1 we can build a decision tree which is found in Fig. 1 to find the possible outcome.

TABLE 1: CLASSIFICATION OF SYSTEMS BASED ON THEIR VULNERABILITIES TO VIRUSES.

System Name	Nimda	Melissa	MyDoom	Storm Worm
A	Yes	No	No	No
B	No	Yes	No	Yes
C	Yes	No	Yes	Yes
D	Yes	No	No	No
E	No	Yes	No	No

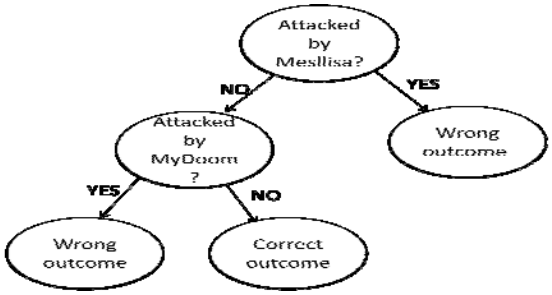

Figure 1: Decision tree for Table1

2.2 Clustering

Clustering is very similar to classification except that the groups are not predefined, but rather are defined by the data alone. The concept of clustering has been around for a long time. It has several applications, particularly in the context of information retrieval and in organizing web resources [4]. The main purpose of clustering is to locate information and in the present day context, to locate most relevant data. For example we have three sets (MIXED) of different colors (say red, blue and yellow) found in fig. 2. After clustering, the sets are grouped separately based on the color as each group will have a unique feature in it, which is found in fig. 3.

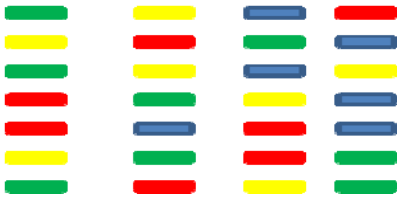

Figure 2: Item sets before clustering

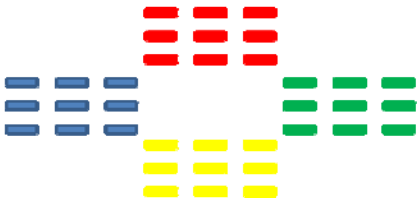

Figure 3: Grouped items after clustering

2.3 Association

Association refers to the data mining task of uncovering relationships among data. As noted earlier, a huge amount of data is being processed my most enterprises every day. A recent survey pointed out that the amount of data processed in the Wal-Mart is almost above 20 million per day. Given this humongous amount of data, it is always better to find and analyze the relationship between the various transactions to arrive at useful information [5]. For example consider a scenario where there are four items (cake, ice-cream, juice and soup) and four transactions (T1, T2, T3, T4). We use 1 to represent that the presence of the item in the transaction and 0 to represent the absence of the item in the transaction. From the data given in the table 2 we can arrive at useful and relevant relationship between the various items using the Association Rule analysis. From the data in the table 2 we can arrive at a conclusion that if the customer buys a Cake and a Ice-Cream he also gets a juice along with it. We can represent it using the notation: {Cake, Ice-Cream} → {Juice}

TABLE 2: DIFFERENT TRANSACTIONAL VALUE

Transactions	Cake	Ice-Cream	Juice	Soup
T1	1	0	0	1
T2	1	1	1	0
T3	0	1	0	1
T4	1	1	1	0

III. A taxonomy of Web Mining

Web mining is the application of machine learning (data mining) techniques to web-based data for the purpose of learning or extracting knowledge. Web mining encompasses wide variety of techniques for finding the relevant information. Web mining methodologies can generally be classified into one of three distinct categories: web usage mining, web structure mining, and web content mining. The graphical representation of the taxonomy of web mining is represented in figure4. Before we deal with the Web usage mining which is our prime focus, we would like to give a brief introduction about the other two types of Web Mining which are web content mining and web structure mining.

A. Web Content Mining

Web content mining is used to discover useful and relevant information from the content of the web pages. This might be very similar to the usage of keywords in a search engine, but it is far useful and efficient than it [6]. As noted earlier, web pages are a source of very rich information which comprises of text, images, audio, video, hyperlinks and metadata. Web content mining mainly focuses on the content of the web pages rather than the links and hyperlinks of that web page.

B. Web Structure Mining

Web Structure Mining deals with the link structure of the web. Web usage mining is the process of extracting useful information from

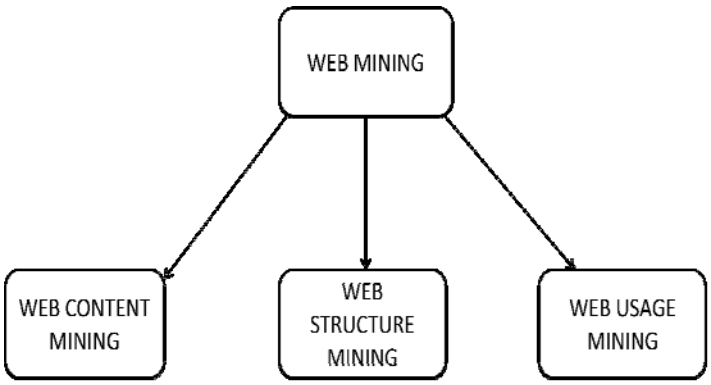

Figure 4: Taxonomy of web mining

server logs i.e users 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. This method is used to find the similarity and the relationship between web pages [7]. Web structure mining focuses on the hyperlink structure of the Web. The different objects are linked in some way. Simply applying the traditional processes and assuming that the events are independent can lead to wrong conclusions. However, the appropriate handling of the links could lead to potential correlations, and then improve the predictive accuracy of the learned models. Two algorithms that have been proposed to lead with those potential correlations and those algorithms are PageRank and the HITS (Hyperlink-Induced Topic Search) [8].

C. Web Usage Mining

Web Usage Mining performs mining on the Web Usage Data, or Web Logs (plural form of a web blog). A Web log is a listing of page reference data [9]. These logs can be examined from a server or client perspective. When evaluated from a server perspective, mining uncovers information about the sites where the server resides. By evaluating a client perspective information about a user (or group of users) can be gathered [10]. For example- The webmaster at the ABC Corporation learns that a high percentage of users have the following pattern of references to the pages:{A,B,A,D}.This means that a user accesses a page A, then page B, then back to page A, and finally to page D. Based on this observation, the webmaster can make a conclusion that a link is directly needed from page B to page D. He adds this link. Web usage mining can also be useful in developing a separate profile for every user, thus aiding in personalization. Using the techniques of Web usage mining, we can keep track of the previously accessed pages which can be used to identify the browsing behavior of a user and by doing this we can achieve Personalization for that user. By gathering information about the browsing behavior of a user, needed links can be identified to improve the overall performance of the web site.

IV. Background Work

Before we deal with the crux of this research paper, we need to have a look at the concepts which are required for the PALM algorithm. In this section we will deal with the Association Analysis in detail which is the basic stepping stone of our research.

A. Association Analysis

This is used for discovering relevant and useful relations between variables in large databases. Association Analysis is frequently used in retail stores, to assist in marketing, advertising and inventory control. They are used to show the relationship existing between data items. Before dealing with the association analysis let us emphasize on the basic concepts of Association Analysis such as Support, Large item set and Candidate set.

B. Definition of Support

Equation(1) defines Support(S) for an item as the ratio of the number of transactions in which the item exists as a component to the total number of transactions present in the database or (2) defines a simpler version of Support(S) which defines support of an item as the number of occurrences the item occurs in all transactions [11].

$$\text{Support(S)} = \{\{\text{number of transactions in which the item is present}) / \text{total number of transactions}\}*100 \quad (1)$$

$$\text{Support(S)} = \{\text{number of occurences in which the item is present in all transactions}\} \quad (2)$$

C. Definition of Large Itemset

When the frequency of occurrences of a specific pattern is above its threshold S, then we call that itemset a Large Itemset(L).

V. Apriori algorithm

Before dealing with the research work, we would maneuver our focus to the general Apriori algorithm which has been already proposed. The Pre-Processed Apriori Algorithm differs from the conventional Apriori Algorithm as it Preprocesse the complete transaction of a database before we find the Large Item set(L) and Candidate set(C) [12]. The Apriori Algorithm is the most well know Association Rule Algorithm and it uses the Large Itemset property which states that “Any Large Itemset has a subset which is also large”. To illustrate about the Apriori Algorithm we give an example where the Web Master at the ABC Corporation maintains a list of pages accessed

frequently by the users. The set of frequently accessed pages are <A,B,C,D,E>;<A,B>;<A,C,D>;<A,B,C,D>;<A,C,D,E>;<A,E>;<A,D,E>;<B,C,D>;<A>

TABLE 3: TRANSACTION OF FREQUENTLY ACCESSED PAGES BY THE WEB MASTER

Transaction	Items
T1	A,B,C,D,E
T2	A,B
T3	A,C,D
T4	A,B,C,D
T5	A,C,D,E
T6	A,E
T7	A,D,E
T8	B,C,D
T9	A

The first step would be to find the Support(S) of each item. Here we use the simplified formula of Support [13]. After finding the support of each item, we would find the Minimum Support value from the table. Here in this example we have a minimum Support value (S_{min}) of 4 which is derived from the table4. Based on this minimum support we form Candidate sets of size 2 having $S_{min} \geq 4$. After forming the Candidate sets we again find the S_{min} value from the table and in this case we find the S_{min} value to be 4 again from table5. Again, based on the minimum support we find itemsets of size 3. The legitimate condition for combining two itemsets of size 2 is "both the itemsets should have a common item between them". Now we find itemsets of size 3.The only item set of size 3 and $S_{min} \geq 4$ is <A,C,D>.Hence we find that the best Pattern to be followed to make sure that the Web Site is very efficient in linking the pages [14].

VI. RESEARCH WORK

Before we deal with the PALM-Algorithm, a note on Map Reducing Algorithm is introduced to make our point clear as the PALM-Algorithm makes use of the Map Reducing Algorithm to find the initial Itemset and Candidate set.

TABLE 4: SUPPORT VALUE FOR EACH ITEM

Item	Support
A	8
B	4
C	5
D	6
E	7

TABLE 5: SUPPORT VALUE FOR TRANSACTIONS OF SIZE 2

Transaction	support
<A,C>	4
<A,D>	5
<A,E>	4
<C,D>	5

A. Map-Reduce Algorithm

Map Reducing Algorithm is a framework introduced by Google in 2004 to support distributed computing on large data sets on clusters of computers [15]. The map reduce algorithm has made it easier for huge organization to work with large data sets as it follow a hierarchical method for breaking down the large data sets into smaller sets to obtain useful information. In the algorithm (PALM) we use the Map Reduce algorithm for the same functionality as we deal with a large data set and we need to break the large data sets into smaller ones. The map reduce algorithm is used to cluster the entire item sets in the database so that the Candidate and the Large item set can be determined by it. Map Reduce is a framework for processing complex and highly sophisticated transaction across huge datasets using a large number of computers (nodes), collectively referred to as a cluster (if all nodes use the same hardware) or a grid (if the nodes use different hardware).Computational processing can occur on data stored either in a file system (unstructured) or in a database(structured). It consists of two steps:"Map" step: The master node takes the input, partitions it up into smaller sub-problems, and distributes them to worker nodes. A worker node may do this again in turn, leading to a multi-level tree structure. The worker node processes the smaller problem, and passes the answer back to its master node."Reduce" step: The master node then collects the

answers to all the sub-problems and combines them in some way to form the output – the answer to the problem it was originally trying to solve [16]. Map Reduce algorithm is used for various data analytic processing as it works well with large data sets. The Map Reduce algorithm is compatible with most of the RDBMS which has large data sets. The Map Reducing algorithm is the first step used in our PALM algorithm. It is used to sort the item sets, eliminate the items which are below the minimum threshold. We find the initial Candidate and Item Set. The map reducing algorithm is divided into two steps. The first step involves a Preprocess algorithm which defines the minimum support required and we invoke the itemset_generation() function [17] which generates the initial Itemset and Candidate set for further scans. The pseudo-code for the Preprocess algorithm and the itemset_generation() are found in fig. 5 and fig. 6 respectively. Thus we have created the initial Candidate Set(C) and the Large Itemset(L) for the transaction present in the database. Now we have to implement the PALM (Preprocessed Apriori For Logical Matching) algorithm for the generated itemset to get the itemsets for further scans.

```
1   PREPROCESS()
2   {
3     N ←No of occurrences of an item
4     While(transaction alive)
5     {
6      Min_support←large number;
7        Map reduce(Candidate Set)
8        {
9           Foreach(transaction)
10           /*map step*/
11             key←<item(key),value>;
12           /*intermediate step*/
13             Store <item,1> in buffer;
14           /*reduce step*/
15             Support←N*100;
16               If(Min_support>Support(item))
17                 Min_support←Support(item);
18    ITEMSET_GENERATION(Min_support);
19    }
```

Figure 5: Preprocess Algorithm

```
1   ITEMSET_GENERATION(Min_support)
2   {
3     Itemset=Ø;
4     size←size of candidate set generated
5           using PREPROCESS();
6     Threshold←Min_support;
7       For(i←1; i≤size; i++)
8       {
9         If(Support[ITEM_I ]>Min_Support)
10         /*Store itemset in database*/
11          Itemset={itemset ∪ ITEM_I };
12      }
13    }
```

Figure 6: Itemset Generation

B. PALM(PREPROCESSED APRIORI FOR LOGICAL MATCHING GENERATION)

The PALM algorithm scans the entire database of transactions and creates Candidate Set(C) and ItemSet(L) for each level of scan. The PALM Algorithm is used to find the best Item Set which we can chose to improve the efficiency of linking pages when we implement it in a Web Site. It can also be used in business to find the best pattern especially in Market based business where we need to choose the products that are to be sold to make a large amount of profit. PALM uses breadth-first search and a tree structure to count candidate item sets efficiently. It generates candidate item sets of length k from item sets of length k − 1. Then it prunes the candidates which have an infrequent sub pattern. According to the downward closure lemma, the candidate set contains all frequent k-length item sets. After that, it scans the transaction database to determine frequent item sets among the candidates. Explanation of PALM Algorithm-The PALM Algorithm uses a counter to count the scans. It gets the initial CANDIDATE SET(C) and ITEMSET(L) from the itemset_generation() function. The algorithm uses a iterative function called AP() to generate the Itemset for the further scans. This iterative function is invoked until the Itemset and Candidate set becomes NULL. The iterative function increments the scan Value every time when the function is called. We initially assume both the Candidate and Itemset are empty. And when itemsets for scan 2 have been generated, then for each item in the itemset is combined with every other items in that item set and combination are generated when both the items considered are not equal. We call the map reduce algorithm for generating the next scan of itemset.

PALM ALGORITHM:

```
1   Input:
2     K←1;
3     Lk←previous Itemset
4     Ck←previous Candidate set
5
6   Begin:
7     While((Lk≠Ø)&&(Ck≠Ø))
8     {
9       K←K+1;
10      Lk←NULL;
11      Ck←NULL;
12        For each item(I) in Lk-1
13        {
14            For each item(J) in Lk-1 where J≠I
15            {
16              Ck←CkU(IUJ);
17            }
18    }
19  //candidate set Ck is generated
20  MAPREDUCE(Ck);
21  }
22  End
```

Figure 7: PALM Algorithm

Thus for generating the Best ITEMSET for a large database or for PATTERN MATCHING for a web page we need to follow the given steps: We Follow the PREPROCESS ALGORITHM for creating and filtering the individual items based on the minimum threshold support .We create the item set using the itemset_generation pseudo code to get the initial Candidate Set and the ItemSet. We Follow the PALM ALGORITHM to get the best Itemset for the given input. The complete architecture of the PALM algorithm is shown in fig. 8. The framework of our work is shown in the fig. 9.

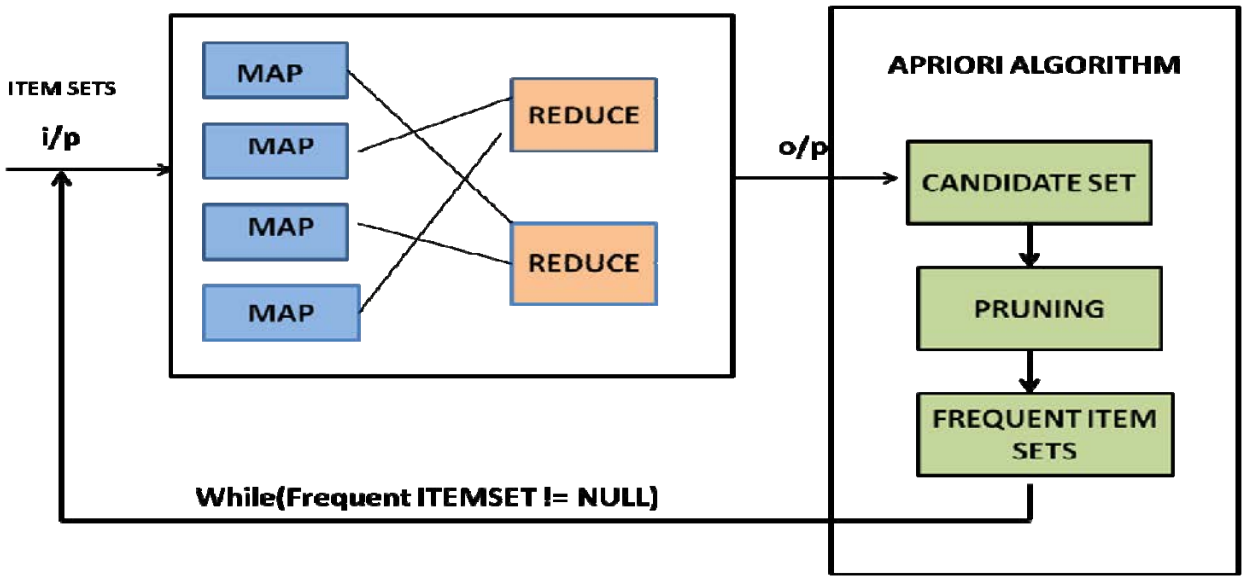

Figure 8: PALM algorithm's Architecture

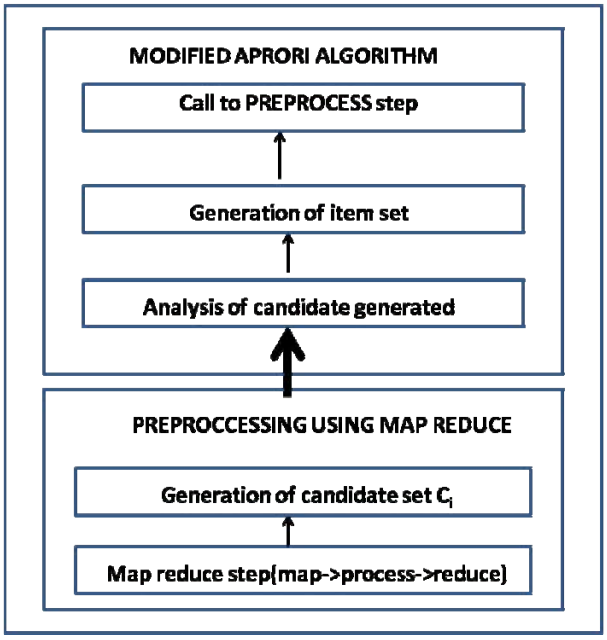

Figure 9: Framework for PALM

C. Experimental Results

The performance of this algorithm for large data sets can be compared with the Apriori Algorithm and we infer that the efficiency of PALM algorithm is better than the Apriori algorithm as it reduces the number of transaction to produce the final itemset. We performed three series experiment with a large data set which consists of 100 records and we implemented both the PALM and the Apriori algorithm. The experimental results are shown in

fig. 9. We can infer from the experiment that the number of transactions (in y-axis) have decreased in the PALM algorithm making it more efficient than the Apriori algorithm.

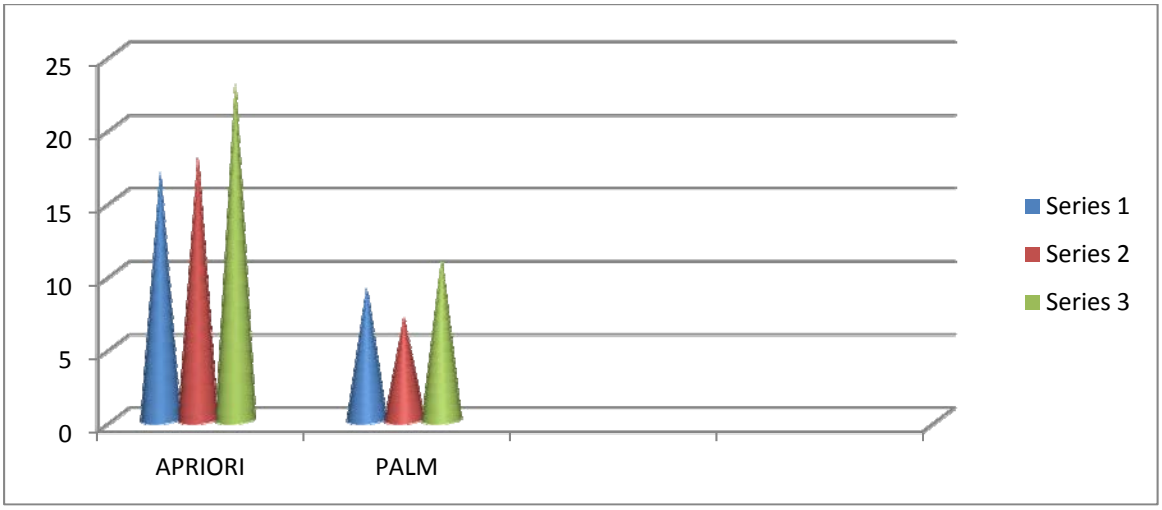

Figure 10: Experimental Results Showing The Efficiency Of PALM

VII. Conclusion and Summary

In this paper we survey the research area of Web mining, focusing on the category of Web usage mining having introduced Web mining. Later in the paper when we had discussed about the Pre-Processed Apriori Algorithm, we reviewed three new algorithms to have an idea about their application and effectiveness. The goal of PALM Algorithm is to increase the efficiency in delivering the content of the Web Pages so as to make it user friendly. The thrust of this research work has been in finding a unique and an efficient algorithm for discovering patterns and we have named the algorithm as PALM (PREPROCESSED APRIORI FOR LOGICAL MATCHING) Algorithm as it is an extension of the generalized Apriori Algorithm. Since this is a huge area, and there a lot of work to do, we hope this paper could be a useful starting point for identifying opportunities for further research.

REFERENCES

[1] Jiawei Han Kevin Chen-Chuan and Chang "Data Mining for Web Intelligence" January 2000.
[2] R. Cooley, B. Mobasher, and J. Srivastava "Web Mining: Information and Pattern Discovery on the World Wide Web"
[3] Thair Nu Phyu “Survey of Classification Techniques in Data Mining” IMECS 2009, March 18 - 20, 2009, Hong Kong.
[4] Raymond T. Ng and Jiawei Han “Efficient and effective clustering methods for Spatial Data Mining” VLDB Chile 1994.
[5] RAJESH NATARAJAN1 and B SHEKAR2 “Interestingness of association rules in data mining: Issues relevant to e-commerce” April/June 2005.
[6] Raymond Kosala and Hendrik Blockeel “Web Mining Research: A Survey” ACM SIGKDD, July 2000.
[7] Shohreh Ajoudanian, and Mohammad Davarpanah Jazi “Deep Web Content Mining” World Academy of Science, Engineering and Technology 49 2009.
[8] Bing Liu “Web Content Mining” Department of Computer Science University of Illinois at Chicago.
[9] Johannes F¨urnkranz “Web Structure Mining”Exploiting the Graph Structure of the World-Wide Web”.
[10] Miguel Gomes da Costa Júnior Zhiguo Gong “Web Structure Mining: An Introduction” Proceedings of the 2005 IEEE International Conference on Information AcquisitionJune 27 - July 3, 2005, Hong Kong and Macau, China.
[11] S.Veeramalai , N.Jaisankar and A.Kannan “Efficient Web Log Mining Using Enhanced Apriori Algorithm with Hash Tree and Fuzzy” Anna University Chennai-600025, Tamil Nadu, India.
[12] Ashish Jain “Apriori Algorithm Implementation” January 29, 2009.
[13] A. Rungsawang, A. Tangpong, P. Laohawee, T. Khampachua “Novel Query Expansion Technique using Apriori Algorithm” Kasetsart University, Bangkok, Thailand.
[14] Huan Wu, Zhigang Lu, Lin Pan, Rongsheng Xu and Wenbao Jiang “An Improved Apriori-based Algorithm for Association Rules Mining” China.
[15] Jimmy Lin and Chris Dyer “Data-Intensive Text Processing with MapReduce” University of Maryland, College Park April 11, 2010.
[16] Ariel Cary, Zhengguo Sun, Vagelis Hristidis, Naphtali Rishe “Experiences on Processing Spatial Data with MapReduce”.
[17] Spyros Blanas, Jignesh M. Patel “A Comparison of Join Algorithms for Log Processing in MapReduce” SIGMOD’10, June 6–11, 2010.