INTERNATIONAL JOURNAL OF TECHNOLOGY AND COMPUTING (IJTC)
ISSN-2455-099X,
Volume 2, Issue 4 April 2016.

PERFORMANCE ENHANCEMENT OF RECLAT ALGORITHM USING MAP REDUCE TECHNIQUE

Kaveri Maniktala [a], Rupinder Kaur Gurm [b]

[a] *Research Scholar, M.Tech CSE*
[a] *Assistant Professor, CSE Department*
[a,b] *Department of CSE,RIMT-IET, Mandigobindgarh , Punjab, India*

ABSTRACT

Data mining is the process of discovering meaningful, novel and interesting patterns from large amount of data. Data mining tools predict future trends and behaviours, allowing businesses to make proactive, knowledge-driven decisions. For discovering relations between variables in large databases association rules are formulated. Association rules are widely used for market basket analysis. Two of the widely used algorithms for generation of association rules are apriority and eclat.

Keywords: market basket analysis, association rules, support, eclat.

I. INTRODUCTION

Data mining is the computer-assisted process of exploring through large amount of data in search of consistent patterns and/or systematic relationships and then extracting useful information. Data for data mining is obtained by warehouse, which is a repository for large amount of data. By using varied techniques to go through large amount of warehoused data, data mining helps various business organizations to recognize significant facts, relationships, trends, patterns, exceptions and anomalies that might otherwise go unnoticed.

Data mining uses the technique of modelling, wherein models, which are either mathematical relationships or set of examples, are built using the cases where situations and the data required for that situation is known and then applying it to cases where answers are not known. This process of applying model to new data is known as scoring.

A. TECHNIQUES IN DATA MINING

The concept of data mining is carried out by using three different techniques.

1. **Prediction Analysis**: Generally in prediction analysis using the historical records of the company, the future trend is predicted. But the original meaning of prediction is to predict the identity of one thing based on another related thing. It is a widely used topic and can be used anywhere to predict the failure of machines or to calculate the company's profitability. Credit card authorizing companies use it to predict fraudulent transactions.

2. **Clustering**: A group of objects that belong to the same class are called clusters. Objects which are similar, based on certain criteria, are part of one cluster and dissimilar objects are a part of another cluster. Clustering, at simple level, is done using one or more attributes as your basis for identifying a cluster of correlating results. Different information can be identified using clustering, as it correlates with other examples so it can be seen where the similarities and ranges agree. Clustering can work in two ways, it can be assumed that there is a cluster at certain point and using the identification criteria it can be seen if it is correct. In another way certain input attributes are given and different artifacts can be identified.

3. **Association:** It is the most straight forward data mining technique. A simple correlation between two or more items is made, often which are of the same type to identify patterns. For example, in a grocery store when a customer buys bread it is most likely that he will buy butter also. These are generally if/then statements used to unravel relationships between seemingly unrelated data from any information repository. Association was initially used in **market basket analysis** to find relation between items purchased by customers. It is now most widely used in catalogue designing, product clustering, store layouts etc.

B. ASSOCIATION RULES

Association rules are if/then statements. If is also known as the antecedent and then as consequent, which is found in combination with antecedent. Association rules are created by analysing the information repository for frequent if/then patterns and using the criteria for support and confidence to identify important relationships.

Support is an indication of how frequently the items appear in the database or information repository. **Confidence** indicates how many times those if/then statements are found to be true.

Formally, let $I=\{I_1, I_2,\ldots\ldots, I_n\}$ be any set of items. Let $DB = \{t_1, t_2, t_3,\ldots\ldots, t_m\}$ be a database of transactions such that each transaction t_i is a set of items and $t_i \subseteq I$. Given, an itemset $x \subseteq I$, a transaction t contains x if and only if $x \subseteq t$. An association rule is of the form X=>Y, where $X \subseteq I$ and $Y \subseteq I$ and $X \cap Y = \emptyset$.

Inorder to select minimum rules from the set of possible rules, constraints of minimum threshold on support and confidence are usually applied. These constraints are user specified. Frequent itemsets are those whose support is greater than or equal to the minimum support threshold. In two steps association rules are formulated. In the first, using minimum support all the frequent itemsets are found in the database. In the second step, using the frequent itemsets and the minimum confidence threshold, rules are formed.

C. ALGORITHMS

There are numbers of algorithm available for finding frequent item sets such as eclat algorithm, apriori algorithm and so on. The main rule of data mining is to discover all the items that have support and confidence greater than or equal to given minimum support and confidence.

Apriori algorithm is the traditional algorithm for the generation of frequent itemsets and hence generates association rules. It uses horizontal datasets. Apriori algorithm works on the apriori principle which states that is an itemset is frequent then all of its subsets must also be frequent. It uses a "bottom-up" approach where in frequent itemsets are generated one at a time, by examining the candidates which are generated, by testing those candidates with the threshold. Firstly, candidate itemsets are generated by using the join operation, then those candidates are checked for minimum support, from this frequent itemsets are generated.

The main problem with algorithm is that the database scan needs to be done every time support has to be calculated. Another issue is candidate generation becomes a bottleneck hence it is very slow. The whole database transactions need to be memory resident.

Eclat algorithm is very simple as it uses vertical layout of database. By using vertical layout, the database scan needs to be done only once. The main operation used in eclat is intersection operation. Only support is calculated in this algorithm, as the support count is the length of the itemset. This algorithm is better than apriori for the generation of frequent itemsets.

The main disadvantage of this algorithm is that if the itemset is too long it takes substantial amount of memory and computation time.

II. PRELIMINARIES

A. Eclat algorithm

It uses bottom-up approach to find all frequent itemsets by using depth first search methodology. It makes use of vertical database layout wherein each item is stored in accordance with it's tidlist and it uses intersection based approach to find the support of an itemset.

TID	Item
1	a,b
2	a,b,d
3	b,c,d,e
4	a,c,e
5	b,c,e

Vertical format:

Itemset	TIDset
A	{1,2,4}
B	{1,2,3,5}
C	{3,4,5}

D	{2,3}
E	{3,4,5}

2-itemsets are generated by using the intersection operation. Here, considering this example, we take minimum support to be 1.

Itemset	TIDset
(a,b)	{1,2}
(a,c)	{4}
(a,d)	{2}
(a,e)	{4}
(b,c)	{3,5}
(b,d)	{2,3}
(b,e)	{3,5}
(c,d)	{3}
(c,e)	{3,4,5}

2-frequent itemset:

Itemset	TIDset
(a,b)	{1,2}
(b,c)	{3,5}
(b,d)	{2,3}
(b,e)	{3,5}
(c,e)	{3,4,5}

3-itemset:

Itemset	TIDset
(a,b,d)	{2}
(a,c,e)	{4}
(b,c,d)	{3}
(b,c,e)	{3,5}
(c,d,e)	{3}

3-frequent itemset:

Itemset	TIDset
(b,c,e)	{3,5}

Eclat algorithm takes less space than apriori if itemsets are small in number. Traditional eclat is suitable for small datasets and requires less time than apriori for the generation of frequent itemsets.

B. MapReduce

MapReduce is inspired by functional programming model and many computational problems can be expressed by using this model. A specific format of key/value pairs is used to describe input and output data. An algorithm is expressed using two functions: map function and reduce function. Both the functions are written by the application developer. The map function iterates overa set of input key/value pairs and generates intermediate result which is the output of map phase as key/value pair. The reduce function iterates over intermediate values and is associated by one key. Then. reduce phase generates zero or more output key/value pairs.

```
map(String key, String value)
// key: document name
// value: document  contents
foreach word a in value:
EmitIntermediate(a,"1");
reduce(String key, Iterator values)
// key: word
// value: list of counts
int result=0;
foreach a in values:
result+=ParseInt(a);
Emit(AsString(result));
```

The map function emits 1 each time "a" is encountered. The reduce function sums up the number of times "a" occurs and gives a final result.

C. MREclat

Using MapReduce with Eclat further increases the efficiency of Eclat algorithm. It is done in three steps:

1. All 2-frequent itemsets are obtained along with their TID lists from the database.

2. Balanced group step, where frequent 1-itemsets are grouped.

3. Parallel mining step, wherein the data which is obtained in the first step is distributed to different computing nodes in accordance with the group it's prefix belongs to.

III. PROBLEM FORMULATION

Eclat algorithm works most efficiently with small datasets, when large datasets are considered the TIDset increases and the intersection opeartion takes a large amount of time and hence its efficiency decreases. When MapReduce is used with Eclat some problem still remains which provides a hindrance to the efficiency of the algorithm. So, the efficiency of the algorithm can further be improved if another algorithm is proposed which provides an enhanced procedure to check all the items from the database. It reduces the number of iterations. Instead of depth first search, breadth first search can be used.

IV. CONCLUSION AND EXPECTED OUTCOMES

Eclat algorithm is used to find the associations rules. Associations rules are frequent itemset which are frequently occur in the database. In previous Eclat algorithm uses vertical form for storing items and it uses bottom up approach. But in previous techniques, accuracy is major issue. So we will introduce new algorithm which will increase accuracy, so by using enhanced Eclat algorithm, we can find frequent items which are helpful for various companies, organizations and online markets. So this enhanced Eclat algorithm provides better result with less time.

REFERENCES

[1] Jiawei Han, Micheline Kamber, and Jian Pei. 2011. *Data Mining: Concepts and Techniques* (3rd ed.). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

[2] Rakesh Agrawal, Tomasz Imielinski, Arun Swami "Mining Association Rules between Sets of Items in Large Databases", Proceedings of the ACM SIGMOD International Conference on Management of Data, ACM, Vol. 22. No. 2, 1993, pp. 207-216.

[3] M. J. Zaki, S. Parthasarathy, M. Ogihara, and W. Li, "New algorithms for fast discovery of association rules," Proceedings of the 3th International Conference Knowledge Discovery and Data Mining, vol.97, 1997, pp. 283-286.

[4] M. Song, S. Rajasekaran,"A Transaction Mapping Algorithm for Frequent Itemsets Mining", IEEE Transactions on Knowledge and Data Engineering,2004.

[5] X. Y. Yang, Z. Liu, and Y. Fu, "MapReduce as a programming model for association rules algorithm on Hadoop," Proceedings of 3rd International Conference on Information Sciences and Interaction Sciences (ICIS). IEEE, pp. 99-102, 2010.

[6] Z. F. Li, X. F. Liu, and X. Cao. "A study on improved Eclat data mining algorithm," Advanced Materials Research, vol. 328, 2011. pp. 1896-1899.

[7] Dr. S. Vijyarani, P. Sathya, "An Efficient Algorithm for Mining Frequent Items in Data Streams", International Journal of Innovative Research in computer and Communication Engineering, vol 1, Issue 3, May 2013, pp 742-747.

[8] Manjit Kaur, Urvashi Garg, "ECLAT Algorithm for Frequent Itemsets Generation ", International Journal of Computer Systems, Volume 01- Issue03, Dec 2014, pp 82-84.

[9] Manalisha Hazarika and Mirzanur Rahman, "MapReduce Based ECLAT Algorithm for Association Rule Mining in Data Mining:MR_ECLAT",International Journal of Computer Science and Engineering, Vol. 3, Issue 1, Jan 2014, 19-28.