ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 2, Issue 2, March 2013

A Novel Approach to Mine Frequent Item sets Of Process Models for Cloud Computing Using Association Rule Mining

Roshani Parate
M.TECH. Computer Science. NRI Institute of Technology, Bhopal (M.P.)
Sitendra Tamarkar
Assistant Professor, NRI Institute of Technology, Bhopal (M.P.)

Abstract*— Process mining provides a new means to improve processes in a variety of application domains These process mining techniques help organizations to uncover their actual business processes Frequent pattern mining algorithms are applied in the dyeing process due to difficulties of doing the coloring process in an efficient way. Previous Work done on dyeing process using apriori algorithm. but apriori algorithm has some drawback like more execution time, can not handle the large amount of data Now, we propose modified apriori algorithm that applied in dyeing process model to generate frequent pattern and remove the drawback of apriori algorithm in term of execution time, handle the large amount of data. Frequent Pattern Mining is most powerful problem in association mining. Most of the algorithms are based on algorithm is a classical algorithm of association rule mining the frequent pattern mining algorithms are applied in the dyeing process due to difficulties of doing the coloring process in an efficient way. Apriori algorithm can significantly reduce mining time by generating pattern candidates that had successfully brought many researchers' attention.

***Index Terms*— Frequent Pattern; Priori; Coloring Process; Pattern Candidates.**

INTRODUCTION

Data mining is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Mining Association rule is a way to find interesting associations among large sets of data items. Using this we have determined the frequent item sets based on a predefined support [6]. By cloud we can say that it is an infrastructure that consists of services delivered through shared datacenters and appearing as a single point of access for consumers' computing needs and also provides demanded resources and/or service over the internet. Sector storage cloud is a distributed storage system that can be deployed over a wide area network and allows users to consume and download large dataset from any location with a high-speed network connection to the system. Sector automatically replicates files for the better reliability access and availability. Sphere compute cloud is a computation service which is built on the top of the sector storage cloud. It allows developers to write certain distributed data intensive parallel applications with several simple Application program interfaces. Data locality is the key factor for the performance in the Sphere. Thus to summarize we can say that sector manages data in form of distributed indexed files, sphere processes that data using sphere processing engine that is applied parallel on every data segment managed by sector Frequent Pattern Mining is most powerful problem in association mining. Most of the algorithms are based on algorithm is a classical algorithm of association rule mining [2,3, 4]. Lots of algorithms for mining association rules and their mutations are proposed on basis of Apriori Algorithm [2, 3]. Most of the previous studies adopt Apriori-like algorithms, which generate-and-test candidates and improving algorithm strategy and structure. Several modifications on apriori algorithm are focused on algorithm Strategy but no one algorithm emphasis on representation of database. A simple approach is if we implement in Transposed database then result is very fast. Recently, different works proposed a new way to mine patterns in transposed databases where a database with thousands of attributes but only tens of objects [2]. In many example attribute are very large than objects or transaction In this case, mining the transposed database runs through a smaller search space. In apriori algorithm each phase is count the support of prune pattern candidate from database. No one algorithm filters or reduces the database in each pass of apriori algorithm to count the support of prune pattern candidate from database. The Apriori algorithm had a major problem of multiple scans through the entire data. It required a lot of space and time.

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 2, Issue 2, March 2013

Apriori Algorithm Scans the database too many times, When the database storing a large number of data services, the limited memory capacity, the system I/O load, considerable time scanning the database will be a very long time, so efficiency is very low. In order to overcome the drawback inherited in Apriori.

The modification in our paper suggests that we do not scan the whole database to count the support for every attribute. This is possible by keeping the count of minimum support and then comparing it with the support of every attribute. The support of an attribute is counted only till the time it reaches the minimum support value. Up to the support for an attribute need not be known. This provision is possible by using a variable named flag in the algorithm. As soon as flag changes its value, the loop is broken and the value for support is noted. In this paper we have discussed an Modified algorithm to mine the data from the cloud using sector/sphere framework with association rules.

Why Is Freq. Pattern Mining Important?

- Discloses an intrinsic and important property of data sets
- Forms the foundation for many essential data mining tasks
  - Association, correlation, and causality analysis
  - Sequential, structural (e.g., sub-graph) patterns
  - Pattern analysis in spatiotemporal, multimedia, time-series, and stream data
  - Classification: associative classification
  - Cluster analysis: frequent pattern-based clustering
  - Broad applications

LITERATURE SURVEY

In this section, we briefly review the most related studies including frequent pattern mining algorithms and parallel and distributed algorithms for frequent pattern mining.

A. Frequent pattern mining algorithms

- The numerous studies on the fast mining of frequent patterns can be classified into following categories.
- **HMine to mine frequent patterns**

In 2001 J. Pei [3] proposed an algorithm called HMine to mine frequent patterns efficiently on a sparse dataset. This algorithm utilizes H-Struct data structure, which has very limited and predictable space overhead, and runs very fast in memory setting, hence modified this algorithm with link structure and reverse order processing

- **Mining frequent item sets without candidate generation**

Han et al. (2000) devised an FP-growth method that mines the complete set of frequent item sets without candidate generation. FP-growth works in a divide-and-conquer way as database size. When database is large, the scalability will decline, even the mining process is interrupted. Kawuu W. Lin et al. proposed a cloud computing Algorithm FD-Mine [9]. The main characteristics of it contain compressing the whole FP-tree for preserving data privacy and abating the network latency, and addressing a better dispatching work set strategy for load balancing than BTPtree [11] which extends TPFP-tree. In system architecture of FD-Mine, there is a kernel

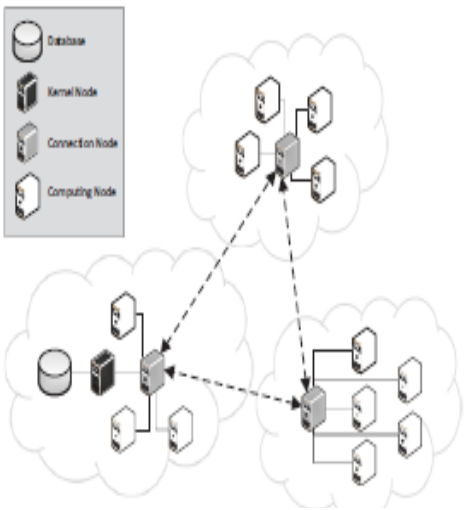


Fig1. Proposed cloud architecture for frequent pattern mining

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 2, Issue 2, March 2013

- **Mining frequent item sets using vertical data format**

Zaki (2000) proposed Equivalence CLASS Transformation (Eclat) algorithm by exploring the vertical data format

- **Mining closed and maximal frequent item sets**

The mining of frequent closed item sets was proposed by Pasquier et al. (1999), where an Apriori -based algorithm called A-Close for such mining was presented

- **Mining high-dimensional datasets and mining colossal patterns**

Pan et al.(2003) proposed CARPENTER , a method for finding closed patterns in high-dimensional biological datasets, which integrates the advantages of vertical data formats and pattern growth methods Pan et al.(2004) proposed COBBLER, to find frequent closed item set by inte-grating row enumeration with column enumeration
The main effects of data mining tools being delivered by the Cloud are:
• The customer only pays for the data mining tools that he needs – that reduces his costs since he doesn't have to pay for complex data mining suites that he is not using exhaustive;
• The customer doesn't have to maintain a hardware infrastructure, as he can apply data mining through a browser – this means that he has to pay only the costs that are generated by using Cloud computing.
Using data mining through Cloud computing reduces the barriers that keep small companies from benefiting of the data mining instruments.
"Cloud Computing denotes the new trend in Internet services that rely on clouds of servers to handle tasks. Data mining in cloud computing is the process of extracting structured information from unstructured or semi-structured web data sources.

- The data mining in Cloud Computing allows organizations to centralize the management of software and data storage, with Apriori principle, apriori algorithm

Agrawal and Srikant (1994) observed an interesting downward closure property, called Apriori, among frequent k -item sets: A k-item set is frequent only if all of its sub-item sets are frequent. This implies that frequent item sets can be mined by first scanning the database to find the frequent 1-itemsets

- **Mining multilevel, multidimensional, and quantitative association rules**

Zhang et al.(2004) considered mining statistical quantitative rules. Statistical quantitative rules are quantitative rules in which the right hand side of a rule can be any statistic that is computed for the segment satisfying the left hand side of the rule.

III PROBLEM IDENTIFICATION

Association rule mining is a popular and well researched area the huge memory space is required for Association Rules mining, so the main memory consumption is usually hard to predict correctly. it take more time for execution Traditional algorithm generates a huge number of candidates for long patterns The efficiency of traditional algorithm is low It is costly to handle a huge number of candidate sets Apriori Algorithm Scans the database too many times When the database storing a large number of data services, the limited memory capacity, the system I/O load, considerable time scanning the database will be a very long time, so efficiency is very low. Slow and provide low accuracy.

IV. PROPOSED APPROCH: IMPROVED APRIORI ALGORITHM

The Apriori algorithm had a major problem of multiple scans through the entire data. It required a lot of space and time. The modification in our paper suggests that we do not scan the whole database to count the support for every attribute. This is possible by keeping the count of minimum support and then comparing it with the support of every attribute. The support of an attribute is counted only till the time it reaches the minimum support value. Up to the support for an attribute need not be known. This provision is possible by using a variable named flag in the algorithm. As soon as flag changes its value, the loop is broken and the value for support is noted. The pseudo code for the proposed algorithm is as follows:
Input: A transposed database D^T and the user defined minimum support threshold s.
Output: The complete set of frequent patterns
Step 1: Convert Database D into transpose form D^T
Step 2: Compute CT_1 candidate transaction sets of size-1 and finds the
Support count.
Step 3: Compute the large transaction sets (*LT*) of size-1.
(i.e., for all *CT1* is greater than or equal to minimum support.)
LT_1 = {Large 1-transaction set (*LT*)};

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 2, Issue 2, March 2013

For (k=2; LT_{k-1} = 0; k++) do
Begin
CTk = Apriori-gen(LT_{k-1},*ct*);
//new candidate transaction sets
End
Return $LT = \cup_k LT_k$;
For all transactions p ∈ LT_{k-1} do begin
For all transactions q ∈ LT_{k-1} do begin
If p.transaction1 = q.transaction,…, p.transactionk-2=q.transactionk-2, p.transactionk-1< q.transactionk-1 then begin
ct=p ∞ q;
If *has_infrequent_subset*(*ct*, LT_{k-1}) then
delete ct;
Else
For all transaction set t∈ D^T do begin
If count(t) <k then delete t;
Else begin
Ct=subset(CT_k, t);
End; End
For all candidate transactions ct∈CT_ido begin
***CT.count* = *CT.count* + 1;**
End; End;
***LTk* = {*ct*∈*CT_k* | *CT.count* ≥ *s*};**
End; End;
End; End;
Return CT_k;

- **Algorithms 3: has_infrequent_subset(ct, LT_{k-1})**
- **For all (k-1)-sub transaction set of ct do**
- **Begin**
- **If t ∈ LT_{k-1} then return true;**
- **else return false;**
- **End.**
- **The main advantage of the proposed algorithm for frequent patterns discovery are, it reduces the size of the database after second pass and, the storage space and saves the computing time**

V. EXPERIMENTAL RESULTS

- To evaluate the efficiency and effectiveness of the improved algorithm, we performed an extensive study of two algorithms: Apriori-like and improved algorithm, on both real time synthetic data sets with different ranges
- Now, we compare the association rules mining algorithms on the whole data set with 5000 data set
- Now we implement the association rules that Apriori algorithm is more efficient which takes less time, less memory and hence results in high efficiency
- The experimental results shows improvement in generation of candidate sets, results in reduced number of data base scan, and also the time and space consumption. we calculate support and confidence

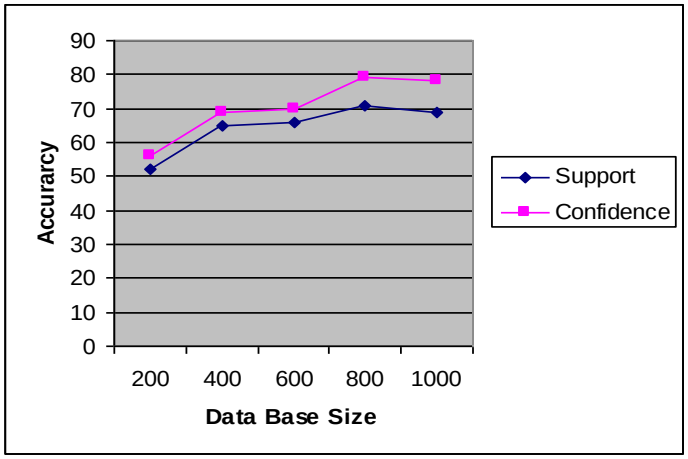
ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 2, Issue 2, March 2013

$$support = \frac{(X \cup Y).count}{n}$$

$$confidence = \frac{(X \cup Y).count}{X.count}$$

Table1: Support and Confidence value of modified Apriori algorithm

Database Size	Modified apriori algorithm	
	Support	Confidence
200	52	56
400	65	69
600	66	70
800	71	79
1000	69	78



VI. CONCLUSION

In this paper we have attempted to give a new perspective algorithm with the eye of a modified apriori algorithm. This algorithm is better than both of the previous methods, i.e., FP Growth tree algorithm and TPFP algorithm. This method works perfectly for data that has been supervised, i.e., data whose classes are already known. But if the classes are not known already, then we can first take any attributes as prominent attributes and test them for modified apriori. Also, the data taken in this example is discrete and this algorithm works on numeric data.

REFERENCES

[1] A new approach for sensitive association rules hiding by Mohammad Naderi ehkordi, Kambiz Badie, Ahma Khade Zadeh in International Journal of Rapid Manufacturing 2009 - Vol. 1, No.2 pp. 128 – 149.

[2] Privacy Preserving Fuzzy Association Rules Hiding in Quantative Data by Manoj Gupta and R C Joshi in International Journal of Computer Theory and Engineering, Vol. 1, No. 4, October, 2009.

[3] Agarwal CC. and Yu PS., "Privacy-preserving data mining: Model and Algorithms, (editors) CharuC.Aggarwal and Philip S. Yu, ISBN: 0-387-70991-8, 2008.

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 2, Issue 2, March 2013

[4] Vassilios S. Verykios,, Ahmed K. Elmagarmid , Elina Bertino, Yucel Saygin, Elena Dasseni. "Association Rule Hiding", IEEE Transactions on knowledge and data engineering, Vol.6, NO.4, April 2004.

[5] Shyue-Liang Wang, Yu-Huei Lee, Billis S., Jafari, A. "Hiding sensitive items in privacy preserving association rule mining", IEEE International Conference on Systems, Man and Cybernetics, Volume 4, 10-13 Oct. 2004 ,Page(s): 3239 – 3244 .

[6] Brin, S., Motwani, R. and Silverstein, C. "Beyond market basket: Generalizing association rules to correlations", in the Proceedings of the 1997 ACM-SIGMOD International Conference on the Management of Data (SIGMOD'97), Tucson, AZ, 1997, pp. 265–276.

[7] Agrawal, R. and Srikant, R. "Mining sequential patterns", in the Proceedings of the 1995 International Conference on the Data Engineering (ICDE'95), 1995. pp. 3–14.

[8] Mannila, H., Toivonen, H. and Verkamo, A.I. Discovery of frequent episodes in event sequences. Data Mining and Knowledge Discovery, 1997, pp. 259–289.