International Journal of Scientific Research in Computer Science and Engineering
Research Paper Vol-1, Issue-3 ISSN: 2320–7639

Mining Frequent Pattern from Large Dynamic Database Using Compacting Data Sets

Nidhi Sethi[1] and Pradeep Sharma[*2]

[1]*Shri vaishnav Institute of Management, Indore, India*
[*2]*Govt. Holkar Science College, Indore, India*

Available online at www.isroset.org

Received: 4 March 2013 Revised: 15 March 2013 Accepted: 08 June 2013 Published: 30 June 2013

Abstract-Frequent pattern mining has been a focused theme in data mining research and the first step in the analysis of data rising in a broad range of applications. Apriori based algorithms have used candidate itemsets generation method, but this approach was highly time-consuming. Several research works have been carried out which can avoid the generating vast volume of candidate itemsets. In this paper, a new approach Compacting Data Sets is introduced. In Compacting Data Sets (CDS) approach first merging of duplicate transactions is being performed and then intersection between itemsets is taken and then deleting unneeded subsets repeatedly. This algorithm differs from all classical frequent itemset discovering algorithms in such a way that it not only removes unnecessary candidate generation but also removes duplicate transactions.

***Keywords* -** Compacting Frequent Pattern Candidate Item sets, Intersection, Duplicate, Unneeded

1. INTRODUCTION

Mining frequent itemsets (or patterns) is a key problem in data mining and it is widely used in applications concerning association rules. A frequent itemset is called maximal frequent itemset (MFI) if it is not a subset of any other frequent itemset. All frequent itemsets are considered implicitly in the maximal frequent itemset because the issue of discovering frequent itemset can be converted to the issue of discovering maximal frequent itemset. Besides, only maximal frequent itemset is needed in some of the data mining applications instead of the frequent itemset. In 1998, Bayardo presented an algorithm of mining maximal frequent itemset denoted as Max-Miner, which has used set-enumeration tree as the concept framework and adopted breadth-first searching method, as well as superset pruning strategy.

2. RELATED WORK

Association mining over dynamic dataset is a challenging area of research for the data mining researchers. Several recent works can be found in the literature to meet this challenge. Some of them are: FUP (Cheung et al., 1996), FUP2 (Cheung et al., 1997), MAAP (Ezeife and Su, 2002), Borders (Feldman et al., 1999), Modified borders (Das and Bhattacharyya, 2005). Next, we report some of these in brief and also discuss their pros and cons.

*MAAP [Ezeife and Su]*This algorithm first finds out the old frequent itemsets that will remain frequent in the updated dataset also. Downward closure property of frequent itemsets makes this job little simpler. Then it checks for possible new frequent itemset, and if found then new candidates are

Corresponding Author: *P Sharma*

FUP [Cheung et al.] FUP first scans the incremental part of the dataset and detects (i) the looser single itemsets, i.e. the itemsets that becomes infrequent due to the inclusion of the incremented part and (ii) it finds the candidate frequent itemsets. Then the whole dataset (i.e. the old and new together) is scanned to find their support in the complete dataset. Next, it performs similar operations iteratively for k-itemsets. Finally, after multiple scanning of the dataset it finds all the maximal frequent sets.

*Borders [Feldman et al., (1999)]*This algorithm finds the frequent itemsets from the dynamic dataset, using the frequent itemsets already discovered from the old dataset. Here, an infrequent itemset is termed as border set if all the non empty proper subsets of it are frequent. Due to the insertion of new records to the dataset, some of the border sets may become frequent, and is termed as promoted border set .For that- the border sets of the old dataset also have to be maintained along with the frequent sets derived. Based on the promoted border set, some new candidate itemsets are generated and checked for frequent set.

3. BASIC CONCEPTS

Let I= {i_1,i_2,...,i_m} be a set of m distinct items. A transaction T is defined as any subset of items in I.
A transaction database D is a set of transactions. A transaction T is said to support an itemset X$\subseteq$I if it contains all items of X. The fraction of the transaction in D that support X is called the support of X, denoted as support(X). An itemset is frequent if its support is above some users

defined minimum support threshold. Otherwise, it is infrequent.

1. The number of items in an itemset is called the length of an itemset. Itemsets of some length k are referred to as k-itemsets.

2. A frequent itemset is called maximal if it is not a subset of any other frequent itemset. .

Property 1: If an itemset is frequent, all its subsets must be frequent.

Property 2: Maximal frequent itemset is a subset of frequent itemsets.

Property 3: if Y is a subset of X (i.e. $Y \subseteq X$), then support(Y) ≥support(X).

4. PROBLEM STATEMENT

Several algorithms have been developed for mining frequent patterns. They are basically divided into two categories:-

(a) Apriori based

(b) FP tree based

Following problems have been observed with these algorithms.

(1) They require huge number of candidate generation

(2) Large number of data scanning is being performed

(3) They work only for static database

(4) Needs lot of input /out operations

5. PROPOSED METHODS

Proposed algorithm is divided into two phases. In the first phase generation of compact dataset for given data base is being carried out, following in the second phase intersection between the rows is performed. Both the phases are being created in order to reduce the length of item set and the volume of data set. Based on length of item sets, first screening of the data set in descending order is done, and then transactions with support count greater then minimal support threshold to a frequent item set are moved in one set, and deletion of all subsets of those transactions to distill the data set is carried.

Step 1: Scan the data set to find frequent one item set.

Step 2: Scan the data set, delete all infrequent 1-itemset from all transactions; and then integrate identical transactions. Then sort the data-set in descending order of length of item set.

Step3: Process every transaction T in S with minimum support count greater then threshold. Move these T in one set denoted and delete all T^3 ($T^3 \subset T^2$ 3>2)

Step 4: Delete all non-MFI.

Step 5: End

	.Updating database	TID	Itemset
D’ Updated database	D Original database	T1	ABC
		T2	ABCF
		T3	ABCE
		T4	ABDF
		T5	CDF
		T6	ABC
		T7	ABCE
		T8	CDE
		19	BDE
	D+	T10	BD
		T11	BDF
		T12	ABCD

Table 1(transactional database)

In the original database there are nine transactions, three more transaction are added to the database .Now total 12 transactions are there in the updated database. For this updated database next we generate compact data set. Suppose 50% minimum support count is taken.

TID	Itemset	Support count	Count with Subset
T1	ABCD	1	1
T2	ABC	5	6
T3	ABD	1	2
T4	BD	3	5
T5	CD	2	3

Table 2 (Compacted data set of D’)

Now to generate frequent item set by using subset count between rows is performed from the first onwards and increment the count of each row.

TID	Itemset	Support count	Sub Set Count
T1	ABC	6	6
T2	ABD	2	2
T3	BD	4	5
T4	CD	3	3

Table 3(Itemset after performing row intersection)

After performing subset count process we get

Frequent three item set {A, B, C}

Frequent two item set {{A, B}, {B, C}, {A, C}}

Frequent One item set {{A}, {B}, {C}, and {D}}

6. PERFORMANCE STUDY

CDS algorithm is easy to implement. It is coded and tested using VB .Net language on a Pentium 3i/2.0 GHz CPU, 1GB memory running Windows 7 operating system. Testing dataset is extracted from a Medical store sales record. There are total 50 different items. We are tested 10000, 15,000 and 20,000 transactions respectively with each record having maximum 10 categories of commodity number. Figure 1 show the running time for different volume of datasets with minimum support threshold of 10%, respectively. The bigger minimum support threshold, the lesser time

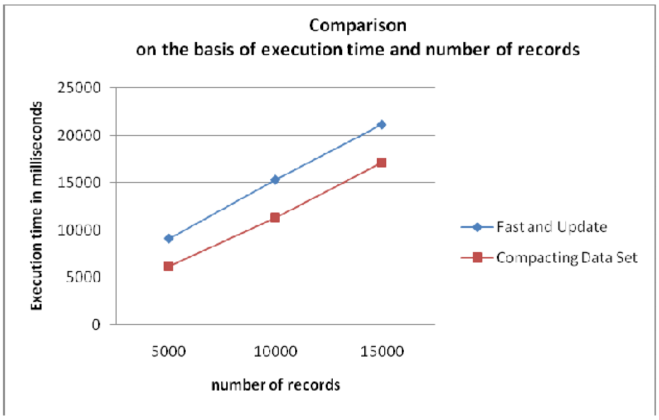

Figure 1

Figure 2 show the running time for 10500 records for different minimum support threshold values i.e.10%, 15% and 20%

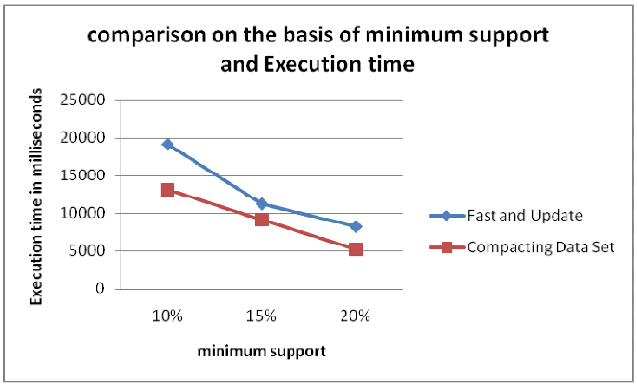

Figure 2

7. CONCLUSION

CDS provides a new and efficient method for discovering frequent pattern; it compact data-set by deleting items in infrequent 1-itemsets and merging duplicate transactions repeatedly, and utilizes the subset of transactions with other transaction itemsets to perform pruning; along with the discovering process, with the increasing of the number of deleted transactions, the amount of time needed for calculating subset will decrease rapidly. It's time & space cost drastically decrease when data-set volume increases, so its usability retains for MFI applications for high volume data-sets. The CDS algorithm can be further optimized in various aspects, such as to keep a record of all resulting sub set to avoid duplicated generation.

REFERENCES

[1] Sotiris Kotsiantis, Dimitris Kanellopoulos Association Rules Mining: A Recent Overview GESTS International Transactions on Computer Science and Engineering, Vol.32 (1), 2006, pp. 71-82

[2] Goswami D.N, Chaturvedi Anshu. Raghuvanshi C.S An Algorithm for Frequent Pattern Mining Based On Apriori Goswami D.N. et. al. / (IJCSE) International Journal on Computer Science and Engineering Vol. 02, No. 04, 2010, 942-947

[3] Ratchadaporn Amornchewin Worapoj Kreesuradej Incremental Association Rule Mining Using Promising Frequent Itemset Algorithm 1-4244-0983-7/07/$25.00 ©2007 IEEE

[4] Sheila A. Abaya Association Rule Mining based on Apriori Algorithm in Minimizing Candidate Generation International Journal of Scientific & Engineering Research Volume 3, Issue 7, July-2012 1 ISSN 2229-5518

[5] Pramod S. O.P. Vyas Survey on Frequent Item set Mining Algorithms International Journal of Computer Applications (0975 - 8887) Volume 1 – No. 15

[6] Dr. R. S . Jadon Dr. R. C. Jain Sunil Joshi An Implementation of Frequent Pattern Mining Algorithm using Dynamic Function International Journal of Computer Applications (0975 – 8887) Volume 9– No.9, November 2010

[7] Pang-Ning Tan Michael Steinbach Vipin Kumar Introduction to Data Mining Copyright c 2006 Pearson Addison-Wesley. All rights reserved

[8] William Cheung and Osmar R. Zaïane Incremental Mining of Frequent Patterns Without Candidate Generation or Support Constraint Proceedings of the Seventh International Database Engineering and Applications Symposium (IDEAS'03) 1098-8068/03 $17.00 © 2003 IEEE

[9] Ravindra Patel, D. K. Swami and K. R. Pardasani Lattice ased Algorithm for Incremental Mining of Association Rules ternational Journal of Theoretical and Applied Computer Sciences Volume 1 Number 1 (2006) pp. 119–128

[10] Deepak Garg, Hemant Sharma Comparative Analysis of Various Approaches Used in Frequent Pattern Mining (IJACSA) International Journal of Advanced Computer Science and Applications, Special Issue on Artificial Intelligence

[11] Endu Duneja A.K. Sachan A Proficient Approach of Incremental Algorithm for Frequent Pattern Mining International Journal of Computer Applications (0975 – 888) Volume 48– No.20, June 2012

[12] B. Nath1, D K Bhattacharyya2 & A Ghosh3 Discovering Association Rules from Incremental Datasets International Journal of Computer Science & CommunicationVol. 1, No. 2, July-December 2010, pp. 433-441

[13] Sandhya Rani Jetti, Sujatha D Mining Frequent Item Sets from incremental database : A single pass approach International Journal of Scientific & Engineering Research, Volume 2, Issue 12, December-2011 1 ISSN 2229-5518

[14] Frequent Item set Mining Methods Jiawei Han und Micheline Kamber. Data Mining – Concepts and Techniques. Chapter 5.2

[15] N L. sarda N V. srinivas an adaptive algorithm for incremental mining of association rules indian institute of

technology bombay. downloaded on april 24, 2009 at 06:04 from ieee xplore.

[16] L Pratima Gautam K. R. Pardasani A Fast Algorithm for Mining Multilevel Association Rule Based on Boolean Matrix Pratima Gautam et. al. / (IJCSE) International Journal on Computer Science and Engineering Vol. 02, No. 03, 2010, 746-752

[17] Stefano Concaro1,2, Lucia Sacchi1, Carlo Cerra2, Pietro Fratino3, and Riccardo Bellazzi1 Mining Healthcare Data with Temporal Association Rules: Improvements and Assessment for a Practical Use C. Combi, Y. Shahar, and A. Abu-Hanna (Eds.): AIME 2009, LNAI 5651, pp. 16–25, 2009. © Springer-Verlag Berlin Heidelberg 2009

[18] Ahmed Taha1, Mohamed Taha1, Hamed Nassar2, Tarek F. Gharib3 DARM: Decremental Association Rules Mining Journal of Intelligent Learning Systems and Applications, 2011, 3, 181-189