Parallel and Distributed Closed Regular Pattern Mining in Large Databases

M.Sreedevi[1], L.S.S.Reddy[2]

[1]School of Computing, K L University, Green fields
Guntur, Andhra Pradesh, India

[2]Department of Computer Science & Engineering, LBRCE, JNTUK
Mylavaram, Andhra Pradesh, India

Abstract

Due to huge increase in the records and dimensions of available databases pattern mining in large databases is a challenging problem. A good number of parallel and distributed FP mining algorithms have been proposed for large and distributed databases based on frequency of item set. Not only the frequency, regularity of item also can be considered as emerging factor in data mining research. Current days closed itemset mining has gained lot of attention in data mining research. So far some algorithms have been developed to mine regular patterns, there is no algorithm exists to mine closed regular patterns in parallel and distributed databases. In this paper we introduce a novel method called PDCRP-method (Parallel and Distributed closed regular pattern) to discover closed regular patterns using vertical data format on large databases. This method works at each local processor which reduces inter processor communication overhead and getting high degree of parallelism generates complete set of closed regular patterns. Our experimental results show that our PDCRP method is highly efficient in large databases.

Keywords: *Regular patterns, Closed regular patterns, Vertical data format, parallel and distributed algorithm, large databases.*

1. Introduction

Now a day mining large databases is a challenging area in data mining and knowledge discovery research. Current literature survey shows that Association rule mining algorithms proposed in data mining which are not sufficient on large databases and still new solutions have to be found. Frequent pattern mining was fundamental and important in data mining research. The Apriori algorithm [1], [2] was the first algorithm to find frequent itemsets based on anti monotone property that was introduced by Agarwal et. al in 1993. Han et.al [3] introduced frequent pattern tree (FP tree) and FP growth algorithm to mine frequent patterns without candidate generation in the year 2000. These FP mining algorithms assume the data as centralized, memory resident and static. However, in real world data mining methods require to handle large databases in which these assumptions are no longer valid. Therefore, researchers focused on large scale parallel and distributed FP mining algorithms [4] to improve scalability and response time. The item is said to be frequent when its occurrence frequency is not less than the user specified minimum support threshold. Occurrence behavior is not sufficient and temporal regularity is also needed in data mining research.

Regular pattern mining is one of thrust areas in data mining research. Tanbeer et.al [5] proposed an algorithm to discover regular patterns based on temporal regularity of pattern in transactional database. The authors constructed a highly compact tree structure RP tree with support descending order and a pattern growth approach to mine regular patterns in static databases. A pattern is said to be regular pattern if its regularity is less than or equal to user specified maximum regularity threshold. The significance of occurrence behavior of item can be considered in a wide range of real world applications. Vijay Kumar et. al [6] proposed an algorithm to mine regular patterns in transactional databases using vertical data format. Closed item set mining gained lot of attention than traditional mining Methods. Closed item set mining is more appropriate than traditional mining process. Wang et.al Proposed a BIDE (Bi-Directional Extension) algorithm to mine closed sequential pattern without candidate key maintenance.

There is some number of parallel and distributed FP mining algorithms [7] which are developed based on Apriori and FP tree algorithms. So their performance also limited by the capabilities of Apriori technique. To the best of our knowledge no algorithm is proposed to mine closed regular patterns in parallel and distributed environment. So in this paper we propose a new method called PDCRP to mine closed regular patterns in parallel and distributed databases with vertical data format on large databases. In this process first we mine regular itemsets and then mine closed regular patterns by considering global maximum regularity and minimum support .Our method mines the local data base using vertical data format to discover all possible closed regular patterns globally with out inter process communication among processors.

The remaining of this paper is organized as follows. Section 2 describes related work section 3 describes problem definition section 4 describes process of mining closed regular pattern mining and section 5 describes experimental results and finally in section6 we conclude this paper.

2. Related Work

Mining regular patterns is one of the thrust area in data mining research. Recently Tanbeer et.al [4] introduced a new problem to mine regular patterns in transactional database which follows regularity of data item in their occurrence behavior. They proposed an algorithm called RP-tree algorithm and the construction process of RP-tree is similar to construction process of FP-tree construction technique, in which RP-tree maintains transaction ids at each node than support count maintenance in FP-tree. In this process it uses two database scans. With the first database scan it creates a header table called regular table which stores its regularity and support values of items. In the second scan RP-tree is constructed based on previously R-table for regular item sets.

Mining closed item sets has been gained lot of attention in present days than traditional frequent mining techniques. Shengnan Cong et.al[8] proposed algorithm called Par-CSP to accomplish parallel mining of sequential patterns on distributed memory systems. In this process they adopt the divide and conquer method to minimize the inter process communication overhead and also use selective sampling technique to address load imbalance problem. Mafruz Zaman et.al [9] proposed an algorithm called ODAM (Optimized Distributed Association Rule Mining) algorithm to minimize communication cost. Tanbeer et.al [10] proposed a novel algorithm based on tree structure called PP-tree (parallel pattern tree) to mine frequent patterns in large databases. Osmar et.al [11] proposed a parallel algorithm MLFPT(multiple local frequent patterns) for parallel mining of frequent patterns based on FP-growth algorithm by having two data base scans by eliminating the candidate items need. MLFPT approach implemented in two stages: In the first stage parallel frequent pattern trees are constructed and mining process will proceed on these constructed trees in the second stage. In this paper we used vertical data format to mine closed regular items in parallel and distributed environment of large databases. The advantage of vertical data format [12] [13] is require only one database scan, it uses simple operations like union, intersection, deletion etc. Non regular items are pruned in this format only.

3. Problem Definition

In this section we describe the concepts of period of item, regular pattern mining, closed regular pattern mining and also define the problem to obtain complete set of closed regular patterns in parallel and distributed environment.

Let $I=\{i_1,i_2,i_3,\ldots.i_m\}$ be a set of items. A set $X = \{ i_1,i_2,i_3,\ldots.i_n \} \subseteq I$, where $j \leq k$ and $j, k \in [1, n]$ is called a pattern or an item set and T = (tid, X) where T is a transaction in database DB, tid is unique transaction identifier and X is a pattern. The size of the database DB is noted as m = |DB|, transaction set T over the database DB is denoted as $T=\{t_1,t_2,t_3,\ldots..t_m\}$.

3.1 Definition 1(Period of X)

Assume t_j^x and t_{j+1}^x are two consecutive transaction in database at one processor. The period of item x can be defined as number of transactions between t_j^x and t_{j+1}^x , $p^x = t_{j+1}^x - t_j^x$. we consider the first transaction is t_{first} which is null transaction i.e $t_{first} =0$ and last transaction is t_{last} which is last transaction of the database at one processor. Period of item X can be defined as the number of times item X appears in different transactions.

3.2 Definition 2(Regularity of X)

if regularity of X is less than or equal to user given regularity threshold then item X is said to be regular itemset.

3.3 Definition 3 (Closed Regular Itemset)

Assume $X = \{x_1,x_2,x_3,\ldots.x_n\}$ be a set of regular itemsets and $Y = \{y_1,y_2,y_3,\ldots.y_n\}$ be other set of regular itemsets, where $X \subseteq Y$ that is X is subset of Y and Y is super set of X, Support count of Y must not be greater than support count of X then X is called closed regular itemset.

3.4 Definition 4 (Parallel and Distributed Closed Regular Pattern)

Assume DS = $\{p_1, p_2, p_3, \ldots\ldots p_4\}$ be a number of partitions in parallel and homogeneous distributed system. The database DB is divided into n number of equal partitions as db_1, db_2, db_3, db_n, and each partition db_i is assigned to each individual processor p_i. Let regularity of item X is represented as $reg_i(X)$ in db_i and support count of item X is represented as $sup_i(X)$ in db_i. We describe $reg_i(X)$ and $sup_i(X)$ as global regularity threshold and global support count of item X in db_i. reg(X) and sup(X) are global regularity threshold and global support count of item X in database DB respectively. Let λ is user given minimum regularity threshold and δ is user given minimum support count. We accumulate all $reg_i(X)$ and $sup_i(X)$ from each processor to find global regularity reg(X) and global support count sup(X) respectively. Closed regular pattern X is mined which satisfies user given global regularity and global minimum support.

4. PDCRP Method

In this section we describe our proposed PDCRP method to mine closed regular patterns in parallel and distributed environment. We consider distributed environment which contains different locations where each location contains resources like processor, memory, etc. Consider the large database and divide this database into small number of partitions equal in size, non-overlapping partitions in order to distribute for n number of processor. We consider the instance database [14] in the process of mining closed regular patterns in the parallel and distributed environment on large databases.

PDCRP method is implemented in two phases. In the first phase regular patterns are mined in parallel based on user given regularity threshold and closed regular patterns are mined from previously mined regular patterns based on user given minimum support threshold.

Phase I:

Input: DB, $\lambda = 5$

Output: Set of regular patterns

Procedure:

1. Let $X_i \subseteq I$ a k-item set at one processor
2. $P_i^x = 0$ for all X_i
3. For each X_i
4. Find the period of X_i
5. $P_i^x = P_{i+1}^x - P_i^x$
6. $reg(X_i)=max(P_i^x)$
7. if $reg(X_i) <= \lambda$
8. X_i is regular itemset
9. Else
10. Delete X_i
11. Repeat the steps 2 to 10 for P_{i+1}^x items.

In phase I we find set of regular patterns at each local processor based on regularity threshold. For example periodicity of itemsets and regular itemsets are shown in table 4 at two local processors.

Table 1 contains database at two processors P_1 and P_2 having nine transaction each. The horizontal database is converted into vertical format at each processor using one database scan which are shown in table 2.In this table the item set < d >is appeared in the transactions<1, 3, 4, 5, 6, 7, 8> at processor P_1 and also appeared in the transactions < 1, 2, 3, 4, 5, 6, 7, 8>at processor P_2. Each processor contains equal number of transactions which are collected from large database every time. So in this method we also maintain the load balancing property. We consider global maximum regularity λ =5 and global minimum support δ = 10, two measures to mine closed regular patterns.

Table 1: Sample database

Tid	Transaction at P1	Transaction at P2
1	a, b, c, d	b, c, d, e
2	a, b	a, c, d
3	a, c, d, e	a, d, e
4	b, d, e	a, b, c, d, e
5	a, c, d, e	a, c, d, e
6	b, c, d	c, d
7	a, d, e	b, c, d
8	a, b, c, d, e	b, d, e
9	a, b, c, e	a, b, c

Table 2: vertical data format.

Itemset	Tids at P1	Tids at P2
a	1, 2, 3, 5, 7, 8, 9	2, 3, 4, 5, 9
b	1, 2, 4, 6, 8,9	1, 4, 7, 8,9
c	1, 3, 5, 6,8,9	1, 2, 4, 5,6, 7, 9
d	1, 3, 4, 5, 6, 7, 8	1, 2, 3, 4, 5, 6, 7,8
e	3, 4, 5, 7, 8, 9	1, 3, 4, 5, 8

Table 3 represents PDCRP header table which simulates regularity and support values of each item set at every processor. This table is somewhat similar to pp-tree [15] header table that collects all local regularity and global support values respectively. our PDCRP method works to mine closed regular patterns in large databases.

Table 3: PDCRP Header table

Items	***P_1***	***P_2***	-----	***P_n***	***Total***
i_1	$reg_1(i_1)$	$reg_2(i_1)$	-----	$reg_n(i_n)$	$max_i(reg_i(i_1))$
	$sup_1(i_1)$	$sup_2(i_1)$	-----	$sup_n(i_n)$	$\Sigma_i sup_i(i_1)$
i_2	$reg_1(i_2)$	$reg_2(i_2)$	-----	$reg_n(i_n)$	$max_i(reg_i(i_2))$
	$sup_1(i_2)$	$sup_2(i_2)$	-----	$sup_n(i_n)$	$\Sigma_i sup_i(i_2)$
¦	¦	¦	----	¦	¦
i_m	$reg_1(i_m)$	$reg_2(i_m)$	-----	$reg_n(i_m)$	$Max_i(reg_i(i_m))$
	$sup_1(i_m)$	$sup_2(i_m)$	-----	$sup_n(i_m)$	$\Sigma_i sup_i(i_m)$

Table 4: PDCRP local tables with p_i^x and reg_i

	Processor P_1		Processor P_2		Max(reg_1x, reg_2x)
Items	P_1^x	Reg_1x	P_2^x	Reg_2x	
a	1, 1, 1, 2, 2, 1, 1	2	2, 1, 1, 1, 4	4	4
b	1, 1, 2, 2, 2, 1	2	1, 3, 3, 1, 1	3	3
c	1, 2, 2, 1, 2, 1	2	1, 1, 2, 1, 1, 1, 2	2	2
d	1, 2, 1, 1, 1, 1, 1, 1	2	1, 1, 1, 1, 1, 1, 1, 1, 1	1	2
e	3, 1, 1, 2, 1, 1	3	1, 2, 1, 1, 3, 1	3	3

Table 4 represents regularity of itemsets at each processor, Itemset < a > is contains their periodicities <1, 1, 1, 2, 2, 1, 1>, its regularity is 2 at processor P_1 and itemset < b > contains their periodicities <2, 1, 1, 1, 4>, its regularity is 4 at processor P_2.consider the maximum regularity of all regularity values at different processors as the regularity of that itemset. The one itemsets < a, b, c, d, e > are regular itemsets based on global regularity threshold $\lambda = 5$.Similarly we find regular two itemsets, three itemsets and so on.

Phase II

Input: DB with Regular item sets, δ=10 (sup-count)

Output: complete set of closed-regular patterns.

1. Let $X_i \subseteq I$ is a regular k-item set
2. Let $X_j \subseteq I$ is regular k + m item set
3. m=1,2,3 ……..n
4. $X_i \subseteq X_j$ for all i <= j
5. Find $Sup(X_i)$, sup-count of X_i
6. Find $Sup(X_j)$, sup-count of X_j
7. If Sup (X_i) > Sup (X_j)
8. X_i is closed-regular item set
9. Else
10. X_i is not closed-regular item set

In phase II we find complete set of closed regular patterns based on user given support count which is considered globally. That is support count of itemset is sum of support counts of itemset at each processor. The support count of itemset is not less than its immediate super set, the item is said to be closed itemset.

IJCSI International Journal of Computer Science Issues, Vol. 10, Issue 2, No 2, March 2013
ISSN (Print): 1694-0814 | ISSN (Online): 1694-0784
www.IJCSI.org

Table 5: One itemsets with local and global supports.

Items	Sup_i at P_1	Sup_i at P_2	$\Sigma(Sup\ P_1, Sup\ P_2)$
a	7	5	12
b	6	5	11
c	6	7	13
d	7	8	15
e	6	5	11

In table 5 the itemsets < a, b, c, d, e > which satisfies both global regularity λ and global minimum support count δ.

Table 6: PDCRP header table with two itemsets

Items	Reg_ix at P_1	Reg_ix at P_2	Max (regs)	Sup_ix at P_1	Sup_ix at P_2	$\Sigma(Sup_i(P_1,P_2)$
~~a,b~~	~~6~~	~~5~~	~~6~~	4	2	~~6~~
a,c	3	4	4	5	4	~~9~~
a,d	2	4	4	5	5	10
a,e	3	4	4	5	3	~~8~~
b,c	5	3	5	4	4	~~8~~
b,d	3	3	3	4	4	~~8~~
b,e	4	4	4	3	3	~~6~~
c,d	2	2	2	6	6	12
c,e	3	4	4	3	3	~~6~~
d,e	3	3	3	5	5	10

Table 6 contains two itemsets with regularity values and support values at each local processor. itemset < a, b > has regularity 6 at P_1, and it has regularity 5 at processor P_2 , consider the maximum of the two regularity values i.e 6 which does not satisfies the minimum regularity threshold value, so itemset < a, b > is not regular itemset. The itemsets < (a, c), (a, d), (a, e),(b, c), (b, d),(b, e),(c, d),(c, e),(d, e) > are regular two itemsets which satisfied minimum regularity threshold considered at global header table.

Regular one Itemsets < a >, < b >, < c >, < d >, < e > are closed regular itemsets, that is support counts of regular one itemsets is greater than the support counts of immediate regular two itemsets which are represented in table 5 and table 6. So itemsets < a, d >, < c, d >, < d, e > which satisfies minimum support threshold value.

The itemsets < a > contains support count of 12 which is greater than the support count 10 of its immediate super set < a, d>. So itemset < a > is closed regular itemset. Similarly itemsets < b >, < c >, < d > and < e > also closed regular one itemsets. Similarly we repeat this process for remaining two itemsets, three itemsets and so on.

In this paper PDCRP header table collects all global regularities and global supports to find closed regular itemsets. However this method works in parallel and distributed environment for large databases which does not require inter processor communication, it requires only one communication at the time of constructing PDCRP header table and also minimizes I/O cost. Hence our proposed PDCRP method is highly efficient for large databases.

5. Experimental Results

In this section we describe the results of our proposed method. We implemented this method from real(Kosarak) and synthetic (T1014D100K) datasets which are usually use in frequent pattern mining http://cvs.buu.ac.th/mining/Datasets/synthesis_data/ and UCI Machine Learning Repository (University of California – Irvine, CA), these are used by Almanden Quest research group to develop frequent patterns in mining process. In our PDCRP method the horizontal database at each local processor is converted into vertical data format. Every processor contains equal number of transactions while mining process is going on. So load balancing among all the processors is also one of factor we considered.

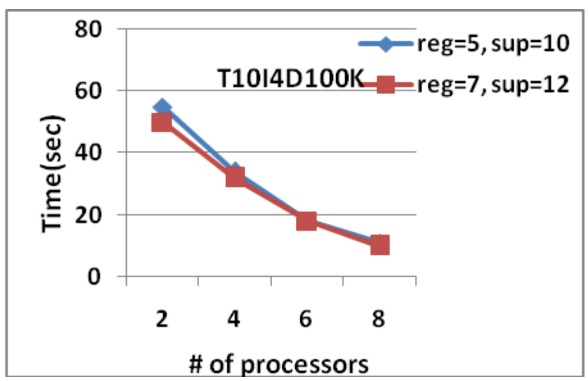

Fig. 1 Execution time on T10I4D100K

We used the systems with 2.66 GHz CPU with 2 GB main memory on Windows XP. We had written programs in java. We distribute the database among processors, so every processor has complete access to its database. We account the results on synthetic dataset T1014D100K which contains 100K transactions, 870 items and its average transaction length is 10.10. W represent execution time for different reg() and sup() values in figure 1.

IJCSI
www.IJCSI.org

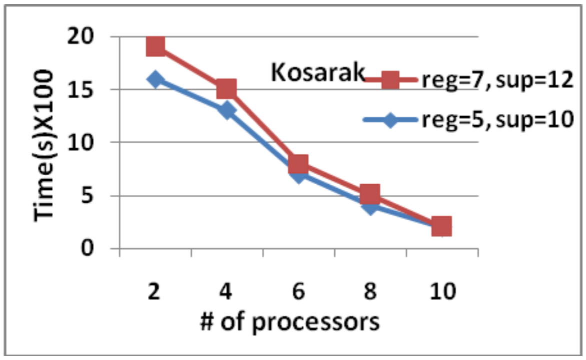

Fig. 2 Execution time on Kosarak

We also account the results on the kosarak real dataset that contains 990 transactions, 41,270 items, and its average length of transaction is 8.10 for different reg() and sup() values which are represented in figure 2.

6. Conclusion

In this paper we introduced a novel approach to mine closed regular patterns using vertical data format in large databases. It takes only one database scan to convert horizontal database to vertical database format. This PDCRP method works in parallel and distributed environment to mine complete set of closed regular patterns based on user given global regularity and support values which minimize I/O cost and no inter processor communication will takes place among processors. So parallel computing is essential component for mining large databases in data mining applications.

References

[1] Agarwal R., Imielinski., Swami N., Association Mining rules between sets of items in large databases. In ACM SIGMOD Int Conference on management of data, 1993, pp.207-216.

[2] Agarwal R and Srikanth R., Fast frequent algorithms for mining association rules. In VLDB, 1994, pp.489-499.

[3] Han J., Pei J., Yin Y., Mining frequent patterns with out candidate generation. In ACM SIGMOD International conference on management of data, 2000, pp.1-12.

[4] Hu J. and Yang Li X. A fast parallel association rules mining algorithm based on FP-forest. In the fifth international symposium on Neural Networks, 2008, pp.40-49.

[5] Tanbeer S K., Farhan C.A., jeong B-S., Lee Y-K. Discovering periodic frequent patterns in transactional data basess. In PAKDD 2009, pp. 242-253.

[6] G.Vijay Kumar, M.Sreedevi, NVS. Pavan Kumar Mining regular patterns in transactional database using vertical data format, In IJARCS,vol 2, No 5,2011, pp.581-583.

[7] Orlando S., Palmerini P., Perego R., Silvestri F.: An efficient parallel and distributed algorithm for counting frequent sets. In VECPAR,2003, pp.242-253.

[8] Shengnan Cong, jiawei Han, David Padua “Parallel Mining of Closed sequential patterns” In ACM KDD, 2005.

[9] mafruz Zaman Ashrafi, David Taniar, and Kate smith. ODAM : An Optimized Distributed Association Rule mining Algorithm. IEEE Distributed systems-IEEE Computer society Vol.5, N0.3, 2004.

[10].Syed Khairuzzaman tanbeer, Chowdhury Farhan Ahmed, Byeong-soo jeong Parallel and distributed algorithms for frequent pattern mining in Large databases. IETE, vol 26, Issue 1, 2009,pp 55-66.

[11] Osmar R.Zaiane., Mohammad El-Hajj., Paul Lu, Fast parallel Association rule mining without candidacy generation., In ICDM ’01 IEEE international conference on data mining ,2001, pp 665-668.

[12] Yi-ming G., and Zhi-jun W., A Vertical format algorithm for mining frequent itemsets. In IEEE transactions, 2010,pp 11-13.

[13] Zaki M.J.; Parallel and distributed association mining: A survey In IEEE concurrency, 1999,pp.14-25.

[14]G.Vijay Kumar., V.Valli Kumari, Parallel and distributed frequent regular pattern mining using vertical format in Large databases, In ARTcom 2012, Banglore,In Press, 2012

[15] Tanbeer S K., Farhan C.A., and jeong B-S., Parallel and Distributed algorithms for frequent pattern mining in large databases. In IETE technical review, vol.26, 2009, pp 55-66.