Closed Regular Pattern Mining Using Vertical Format

M.Sreedevi
Department of CSE
K L University
Guntur, Andhra Pradesh, India.
msreedevi_27@kluniversity.in

L.S.S.Reddy
Department of CSE
LBR College of Engineering
Mylavaram , Andhra Pradesh, India.
director@lbrce.ac.in

Abstract— Discovering interesting patterns in transactional databases is often a challenging area by the length of patterns and number of transactions in data mining, which is prohibitively expensive in both time and space. Closed itemset mining is introduced from traditional frequent pattern mining and having its own importance in data mining applications. Recently, regular itemset mining gained lot of attention in data mining research because of its occurrence behavior. In this paper we propose a new method called CRP-method (closed regular pattern method) to mine closed regular itemsets in transactional database by using vertical data format. Our CRP-method generates complete set of -closed regular patterns in transactional databases for a user given regularity threshold and support. Our experimental results show that this method is efficient in memory and execution time.

Keywords: closed patterns; regular pattern; transactional databases; vertical data format; closed regular patterns;

I. INTRODUCTION

Discovering closed patterns [1], [7], [12] from various domains is a challenging area in data mining and knowledge discovery research. Frequent pattern mining is a traditional, fundamental and essential area in data mining [4], [2], [5], [6]. However the significance of a pattern may not always depend on its frequency (i e., support).Closed pattern mining has number of applications including query access patterns, discovery of DNA sequences, customer shopping sequences, web page sequences and stock-market etc., The significance of a pattern may also depend upon their occurrence characteristics such as occurring at regular intervals in transactional databases. There is no algorithm for mining closed regular patterns in transactional databases. Therefore in this paper we propose a new method called CRP-method (closed regular pattern method) to mine closed regular patterns using vertical data format. In this method there are two phases for mining closed regular patterns. In the first phase regular itemsets are mined based on user given regularity threshold. In the second phase closed itemsets are extracted from previously discovered regular itemsets. The main idea of our new method is to develop a simple, powerful method to mine closed patterns which occur at regular intervals in transactional databases using vertical data format. The experimental results show that the effectiveness of CRP method for finding closed regular patterns in transactional databases.

The remaining of this paper is organized as follows. Section 2 summarizes the closed itemset mining and regular pattern mining. Section 3 describes the problem definition of closed regular pattern mining. Section 4 describes CRP-method to find closed regular patterns using vertical data format in transactional databases. Section 5, Experimental results are shown. We conclude the paper in section 6.

II. RELATED WORK

Frequent itemset mining is one of the important techniques in data mining and it was first introduced by Srikanth and Agarwal for mining frequent itemsets in the year 1993 [5], [6]. It extracts all frequent patterns, correlations, associations among sets of items in transactional databases. The main drawback with this classical algorithm is that it needs repeated scans to generate candidate sets. Han et al., [2] introduced a tree based data structure called FP-tree to generate frequent patterns without generating candidate sets. This algorithm needs only two database scans to mine frequent patterns. Tanbeer et al., [3] introduced a new problem of discovering regular patterns that follow temporal regularity in their occurrence behavior. With the help of regularity measure at which pattern occurs in a database at a user given maximum interval is called regular pattern. They proposed a tree based data structure called RP_Tree, discovers regular patterns in a transactional databases which needs only two database scans. Periodic patterns [11] and Regular patterns [3] are closely related to our work but we

can't directly apply these two algorithms to find closed regular patterns because these two algorithms do not consider the support threshold.

There have been good number of interesting and effective algorithms which are used to mine closed patterns [1], [7], [10], [11], [12]. A Frequent item set I is closed if there exist no super set of I with same support in the database. CLOSpan [7] developed for mining closed sequential patterns in large datasets. CHARM is to mine closed item sets, but is more complicated. BIDE [10] algorithm can only mine frequent closed sequences of single items. But in some cases we need to mine frequent closed sequences of subsets of items, i e., each transaction may contain a set of unordered items. Parallel closed sequential pattern mining (par-csp) algorithm for the closed pattern mining by exploit the divide and conquer property to minimize inter processor communication [8].

III. PROBLEM DEFINITION

Let $I = \{i_1, i_2, i_3, \ldots, i_n\}$ be a set of items. A set $X = \{i_1, i_2, \ldots, i_q\} \subseteq I$, where $l \leq q$ and $l, q \in [1, n]$ is called an item set or a pattern. A transaction $t = (tid, Y)$ is a tuple where tid is a transaction Id and Y is a pattern. The set of transactions $T = \{t_1, t_2, \ldots, t_m\}$ is a transactional database DB over I. $m = |DB|$ is the size of data base.

A. Definition 1 (Regularity of pattern X)

Let t^X_{j+1} and t^X_j, $j \in [1, (m - 1)]$ be two successive transactions where X appears in a DB. The difference between these two successive transactions be able to be defined as a period of X, say p^X (i.e., $p^X = t^X_{j+1} - t^X$, $j \in [1, (m - 1)]$). To calculate the period of a pattern, we consider the first transaction in the *DB* as null i.e., $t_{first} = 0$ and the last transaction is the m^{th} transaction i.e., $t_{last} = t_m$. Let for a T^X, P^X be the set of all periods of X i.e., $P^X = \{p_1^X, \ldots, p_r^X\}$, where r is the total number of periods in P^X. Then the regularity of X can be denoted as $reg(X) = max\{p_1^X, \ldots, p_r^X\}$. A pattern is called regular pattern if its regularity is not more than the user given maximum regularity threshold called max-reg (λ) with $1 <= \lambda <= |DB|$.

B. Definition 2 (closed itemset)

The item set X is closed in a transactional database DB if there exist no proper superset S has same support as X in DB. For example, $\alpha = (a_1, a_2, \ldots, a_k)$ is an itemset and $\beta = (b_1, b_2, \ldots, b_l)$ is another itemset. Let α is a subset of β (i e., $\alpha \subseteq \beta$) therefore β contains α. The support of itemset α is denoted as $Sup(\alpha)$ in database DB i e., number of times the database DB containing α and the support of itemset β is denoted as $Sup(\beta)$ in the given database DB. If $Sup(\beta) !< Sup(\alpha)$ then itemset 'α' is a closed itemset.

C. Definition 3 (closed regular itemset)

X is closed regular itemset in database DB if X is both regular and then closed. The set of closed regular patterns defined as $CRP = \{\alpha \mid \alpha \in RP \text{ and } \beta \in RP\}$ where $\alpha \subseteq \beta$. Regular itemset α is closed regular if $Sup(\beta)$ is not less than $Sup(\alpha)$.

IV. CLOSED REGULAR PATTERN MINING

Closed regular pattern mining is implemented in two phases. First phase, the transactional database DB will convert into vertical data format to mine regular itemsets based on the regularity threshold of an itemsets. In the second phase, closed regular patterns are mined from previously mined regular itemsets

TABLE1 TRANSACTIONAL DATABASE

Tid	Item Sets
1	d, a
2	c, b, a, e
3	b, e, a
4	a, e, b, c
5	a, b, f, e
6	c, d, b
7	c, e, d
8	d, e
9	d, b, c

Consider the transactional database from [3] as our running example to mine closed regular patterns. In the first phase the transactional database (Table 1) is converted into vertical data format (Table 2) i.e., (X, Tid). X is an itemset and Tid is a transaction Id. Table 2 contains item sets and their corresponding transaction ids with their regularity values. The regularity of each itemset is calculated based on the periodicity of itemset which is shown in the Phase I.

Phase I.

Input: Transactional Database (DB), Minimum regularity threshold λ

Output: Set of Regular Patterns

Procedure:

1. Convert Horizontal DB to Vertical DB
2. Let Xi $\subseteq$ I a k-item set
3. $P^X_i = 0$ for all X_i
4. For each X_i
5. Find the period of X_i
6. $P^X_i = P^X_{i+1} - P^X_i$
7. $reg(X_i) = max(P^X_i)$
8. repeat
9. if $reg(X_i) <= \lambda$
10. X_i is regular item set
11. Else
12. Delete X_i .

In phase I the input is transactional database and minimum regularity threshold. The output is regular itemsets. First the horizontal database is converted into vertical format, then transaction difference(p^x) is calculated for each itemset and maximum transaction difference value is considered as regularity of itemset which in steps 5 to 7. Repeat this process for every itemset in DB.Suppose regularity of itemset is less than or equal to minimum regularity threshold (i.e. reg <=λ) then itemset is said to be regular itemset otherwise itemset is not regular itemset (shown in steps 9 to 12).

TABLE 2 VERTICAL DATA FORMAT WITH P^X AND REG VALUES.

Itemsets	**Tids**	**p^x**	**reg**
a	1, 2, 3, 4, 5	1,1,1,1,1,4	4
b	2, 3, 4, 5, 6, 9	2, 1, 1, 1, 1, 3	3
c	2, 4, 6, 7, 9	2, 2, 2, 1, 2	2
d	1, 6, 7, 8, 9	1, 5, 1, 1, 1	5
e	2, 3, 4, 5, 7, 8	2, 1, 1, 1, 2, 1, 1	2
f	5	2, 1, 1, 1, 1, 3	3

Table 2 shows the itemsets and their corresponding transactions where each itemset occurs in the transactions. After converting the database into vertical data format, periodicity i e., P^X is calculated for each item set. For simplicity we assume the first transaction as $t_{first} = t_0$ which is a null transaction and last transaction is $t_{last} = t_n$. Regularity of an item set is obtained from P^X which is maximum periodicity of that item set. Regularity of item is calculated based on transaction difference. For example, item set <a> is appeared in transactions <1, 2, 3, 4, 5>, item set <b> is appeared in transactions <2, 3, 4, 5, 6, 9>. The transaction difference of item set < a > is <1, 1, 1, 1, 1, 4>and the transaction difference of item set < b > is < 2, 1, 1, 1, 1, 3>. Maximum transaction difference is considered as regularity of itemset reg. Let us consider the λ= 4. The itemsets <a, b, c, e, f> are regular itemsets which have been satisfied condition that is reg <= λ.The itemset < d> is not regular itemset because its regularity is greater thanλ.

Phase II

Input: Regular itemsets, sup (support)

Output: Complete set of Closed regular Itemsets

1. Let $X_i \subseteq I$ is a regular k-itemset
2. Let $X_j \subseteq I$ is a regular k+m itemset
3. m= 1, 2, 3, ……, n
4. $X_i \subseteq X_j$ for all i <= j
5. Find Sup(X_i), support-count of X_i
6. Find Sup(X_j), support-count of X_j
7. If Sup(X_i) > Sup(X_j)
8. X_i is closed-regular itemset
9. Else
10. Delete X_i .

In phase II the input is regular itemsets and minimum support threshold (sup) and output is complete set of closed regular itemsets. X_i is set of regular k itemset and X_j is set of regular k+m itemset. Itemset X_i is subset of itemset X_j. Calculate the support counts of k itemset and k+m itemsets. The support count of X_i is greater than support count of X_j then itemset X_i is closed regular itemset other wise X_i is not closed regular itemset.

TABLE 3. ONE ITEMSETS WITH SUP

Itemset	Sup
a	5
b	6
c	5
e	6
f	1

Table 3 contains regular one itemset and their corresponding support values. Similarly table 4 contains regular two itemsets and their corresponding support values. Itemset < a > is subset of itemsets < a, b >, < a, c >, < a, e > and < a, f > and these itemsets are supersets of itemset < a >. Check the support threshold of itemset < a > with their support thresholds of super itemsets < a, b >, < a, c >, < a, e > and < a, f >. The support of < a > is 5 which is not less than support thresholds of < a, b >, < a, c >, < a, e > and < a, f > (i.e. <4, 2, 4, 1>). So itemset < a > is closed regular itemset. Similarly support threshold of regular itemset < b > is not less than support thresholds of regular itemsets < b, c >, <b, e>, <b, f>.So, itemset < b > also closed regular itemset. We continue this process for k+m itemsets.

TABLE 4. TWO ITEMSETS WITH SUP.

Itemset	Sup
a, b	4
a, c	2
a, e	4
a, f	1
b, c	4
b, e	4
b, f	1
c, e	3
c, f	---
e, f	1

V. EXPERIMENTAL RESULTS

We consider the sparse synthetic dataset (T1014D100k) and real dataset (kosarak) which are generally use for experimental analysis for frequent pattern mining as well as regular pattern mining developed at IBM Almaden Quest research group and obtained http://cvs.buu.ac.th/mining/Datasets/synthesis_data/. We did experimental analysis on these data sets with CRP-method with different regularity and support values. All these experiments are done in java on windows XP having the configuration of 2.66 GHz with 2 GB main memory. Our proposed method is to find closed regular patterns in vertical data format. CRP-method is a new approach which mines closed patterns on regular itemsets. We compare our results with RP tree that finds only regular itemset on full dataset of T1014D100K database and kosarak database over different support and regularity thresholds shown in figures 1and 2.

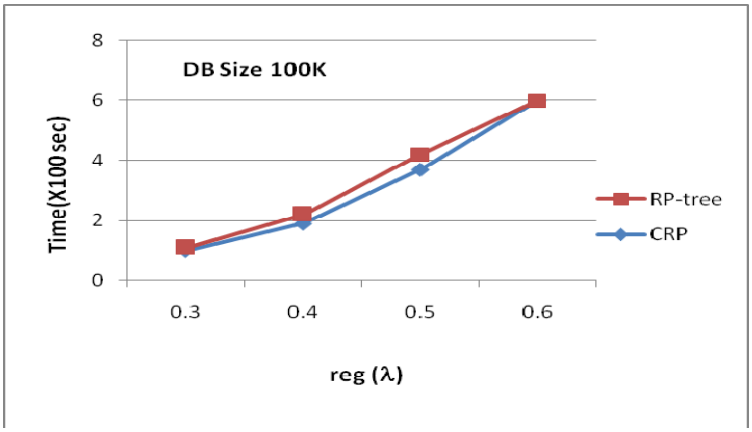

Fig 1. Execution Time over T1014D100K.

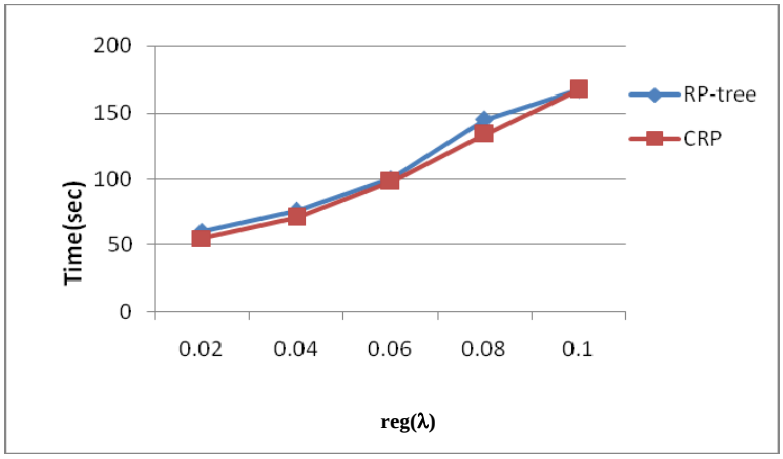

Fig 2. Execution Time over Kosarak.(DB size 500K)

We consider T1014D100K dataset which contains 1,00,759 transactions, 870 items with average transaction length of 10.10 and also with Kosarak dataset which contains 9,90,000 transactions, 41,270 items with average transaction length of 8.10. We produce the results on T1014D100K dataset from 100K and 500K on kosarak dataset which are shown in figures 1 and 2. The higher the max-reg values, the longer the overall time required which are shown by both the methods. However the results clearly demonstrate that CRP method outperforms RP-tree in terms of performance time for different support and regularity threshold values.

VI CONCLUSION

In this paper we proposed an efficient method called CRP-method to mine closed regular patterns in transactional database using vertical data format which is efficient over RP tree in time and space. The advantage of this CRP-method is it uses simple operations like ‘and’ operation, addition, subtraction and simple arrays. Our experimental results show the execution time over different support and regularity threshold values.

REFERENCES

[1] J.Pei, J Han, R Mao "CLOSET" An efficient algorithm for mining frequent closed itemsets" ACM-SIGMOD Int workshop Data mining and knowledge discovery (DMKDOO) pages 11-20 Dallas TX. May 2000.

[2] J. Han, Y. Yin, "Mining frequent patterns with out candidata generation". In Proc. ACM SIGMOD international conf on managt of Data. pages. 1-12 2000.

[3] S.K. Tanbeer, C.F. Ahmed, B.S.jeong and Y.K.Lee "Mining regular patterns in transactional databases." IEICE trans. On Infor systems, pages 2568-2577 2008.

[4] Jiawei Han, Micheline Kamber "Data Mining: concepts and techniques." 2nd ed. An Imprint of Elsevier, Morgan Kaufmann publishers, pages. 232-248 2006.

[5] R. Agarwal, Imielinski, T. swamy A.N. "Mining association rules between sets of items in large data bases' ACM, SIGMOD conference of management of data. pages 207-216 1993.

[6] R. Agarwal., R. srikanth "Fast algorithms for mining association rules" In proc of International conference on very large databases, Santiago, chile, pages 487-499 sept 1994.

[7] Xifeng yan, jiaweihan and Ramin afshar "Clospan: mining closed sequential patterns in large data sets" In proc of int SIAM conf on data mining (SDM\'03) pages 163-177 2003.

[8] Shengnan Cong, Jiawei Han, and David Padua "Parallel Mining of Closed Sequential Patterns" *KDD'05,* August 21.24, Chicago, Illinois, USA 2005.

[9] B.Ozden, S. Ramaswamy, A. Silberschatz "Cyclic Association Rules" 14th International conference on Data Engineering, pags 412-421 2006.

[10] Jianyong wang, jiawei Han and Chun Li "Frequent closed sequence mining without candidate maintenance'. IEEE Transactions on Knowledge and data Engineering, vol. 19, No. 8 August 2007.

[11] M.G. Elfeky, W.G. Aref, A. Elmagarmid " K periodicity detection in time series databases." IEEE Transactiona on Knowledge and Data Engineering 17(7) pages 875-887 2005.

[12] Zaki, M.J.., Hsia, c-J.: Efficient Algorithms for Mining Closed Itemsets and their Lattice Struture. IEEE Trans. on Knowledge and Data Engineering. 17(4), 462-478 (2005).