International Journal of Applied Science and Advance Technology
January-June 2012, Vol. 1, No. 1, pp. 1-3

MINING CLOSED FREQUENT INTERVALS FROM INTERVAL DATA

Anjana K. Mahanta[1] and Mala Dutta[2]

Abstract: Interval data is widely encountered in several areas of data mining applications. In the context of discovering interesting temporal patterns in interval data, the notion of closed frequent intervals is proposed in this paper. A pioneering method to determine the set of closed frequent intervals in an interval database is presented. A rigorous mathematical proof has been provided to substantiate the correctness of the proposed method.

Keywords: Temporal Data Mining, Interval Data, Closed Frequent Intervals

1. INTRODUCTION

Temporal data mining discovers underlying patterns in huge data repositories that are overlooked when the temporal component is ignored. Temporal association rule generation ([1],[2],[3],[4]) sequential pattern mining ([5],[8],[9]) etc. are important fields of research. However most of the work done in this field have focused on time-stamped data i.e. data interpreted as time-points. Interval data and knowledge discovery from databases containing intervals has been largely overlooked. Interval data is however encountered in several data mining application areas. Temporal data stored in data archives have certain attributes. Sometimes a data attribute is itself an interval. Native interval data describing ranges of variable values, for instance, daily stock prices, daily temperatures etc. are also widely encountered. The discovery of underlying patterns and extracting useful information from repositories of interval data has gathered significant momentum in the last few years and is now an active field of research. [10] provides a formal basis for characterizing temporal data models and precisely defines the notions of point-based and interval-based temporal data models. [11] discusses rule semantics in a sequence of interval data. [12], [13] propose methods for extracting patterns from interval data. [15] presents a technique to extract minimal infrequent multi-dimensional intervals. [6] proposes a method to mine maximal frequent intervals. [7] improves some aspects of the method proposed in [6]. [14] gives a method to discover maximal dense intervals from temporal interval data. [16], [17] present techniques to discover closed frequent itemsets. But the problem of generating closed frequent itemsets is NP-hard. In this paper, we propose the notion of closed frequent intervals. Instead of using the NP-hard techniques to find closed frequent intervals, the efficient method proposed in this paper can be used to obtain the closed frequent intervals in a database of intervals. Closed frequent intervals are superior to maximal frequent intervals (see Section 4.1). The technique proposed in this paper uses the set of maximal frequent intervals in an interval database to effectively discover the closed frequent intervals in that database. The correctness of the proposed method to mine closed frequent intervals is established mathematically.

[1] Department of Computer Science, Gauhati University, Assam, India, *E-mail: anjanagu@yahoo.co.in*
[2] Department of Computer Science, Gauhati University, Assam, India, *E-mail: maladuttasid@gmail.com*

2. PRELIMINARIES

Transaction: An ordered pair (tid, I) is called a transaction; tid is a transaction identifier and I is an interval. For a transaction T = (tid, I), intv(T) denotes the interval I.

Transaction database: A transaction database TD is a collection of transactions.

Support of an interval: The support of an interval J in a transaction database TD, denoted by support (J, TD), is the number of transactions T in TD in which J intv(T).

Frequent Interval: Given a minimum support threshold k, an interval I is called frequent in a transaction database TD, if support (I, TD) $\geq k$. The set of frequent intervals in a transaction database TD is denoted by $F(TD)$.

Maximal Frequent Interval: An interval I is called a maximal frequent interval in a transaction database TD if I is a frequent interval and there is no other frequent interval I' in TD such that $I \subset I'$. The set of maximal frequent intervals in a transaction database TD is denoted by $M(TD)$.

Closed Frequent Interval: An interval I is called a closed frequent interval in a transaction database TD if I is a frequent interval and there is no other interval I' in TD such that $I \subset I'$ and support (I, TD) = support (I', TD). The set of closed frequent intervals in a transaction database TD is denoted by $C(TD)$.

3. PROBLEM DEFINITION

Let TD be a transaction database. Let k be the minimum support threshold; k may be expressed in terms of absolute

number of transactions or may be expressed as a percentage of the total number of transactions. The problem of mining closed frequent intervals is to determine $C(TD)$.

4. CLOSED FREQUENT INTERVALS

4.1 Closed Frequent Intervals Vs. Maximal Frequent Intervals

The set of maximal frequent intervals in a transaction database TD gives the most compact representation of all the frequent intervals in the database because if $J \in F(TD)$, then there is some $K \in M(TD)$ such that $J \subseteq K$. Maximal frequent intervals however do not contain the support information of their subsets. An additional pass over the transaction database is required to determine the support of the non-maximal frequent intervals. However on the other hand, the set of closed frequent intervals not only gives a relatively compact representation of all the frequent intervals in the database but also retains the support information of all the non-closed frequent intervals. To determine the support information of all non-closed frequent intervals, an additional pass over the transaction database is not required. For this reason, closed frequent intervals are superior to both frequent intervals and maximal frequent intervals.

4.2 Extracting Closed Frequent Intervals

Let Maxfreq(k) denote the set of all maximal frequent intervals where k is the minimum support value. The following algorithm is proposed to extract closed frequent intervals in a transaction database containing n intervals, where min_sup is the minimum support threshold:

```
1. begin
2.   Cf = NIL
3.    for k = n downto minsup
4.      for every I in Maxfreq(k)
5.        if I is not in Cf then insert (I, k) in Cf
6 .     end for
7.    end for
8. end
```

Cf gives the required set closed frequent intervals along with their respective support values.

4.3 Correctness Claims

Let closedfreq(k) denote the set of all frequent intervals where k is the minimum support value and let maxfreq(k) denote the set of all maximal frequent intervals where again k is the minimum support value.

Then, $\text{closedfreq(minsup)} = \bigcup_{k=\min\sup}^{n} \text{maxfreq}(k)$ where n is the size of the dataset.

Proof: Let $S \in \bigcup_{k=\min\sup}^{n} \text{maxfreq}(k)$

Then $s \in \text{maxfreq}(k')$ for some k' where $\text{minsup} \le k' \le n$

$\Rightarrow s \in \text{closedfreq}(k')$ since every maximal interval is closed.

$\Rightarrow s \in \text{closedfreq(minsup)}$ since $\text{minsup} \le k'$

$\Rightarrow \bigcup_{k=\min\sup}^{n} \text{maxfreq}(k) \subseteq \text{closedfreq(minsup)}$

Next suppose that $s \in$ closedfreq(minsup)

$\Rightarrow s$ is closed and support of s is greater than or equal to minsup

Suppose support of s is k'

Since s is closed, no superset of s will have support equal to k' and so obviously $s \in \text{maxfreq}(k')$

Again, since $n \ge k' \ge \text{minsup}$, $S \in \bigcup_{k=\min\sup}^{n} \text{maxfreq}(k)$

Therefore, $\text{closedfreq(minsup)} \subseteq \bigcup_{k=\min\sup}^{n} \text{maxfreq}(k)$

Therefore, $\text{closedfreq(minsup)} = \bigcup_{k=\min\sup}^{n} \text{maxfreq}(k)$

Hence proved.

4.4 An Example

Consider the following transaction database (Table 1). Here, the number of intervals n = 11 and min_sup = 4. The intervals generated by Maxfreq(k) and the intervals added to *Cf* in each iteration of the for-loop (statements 3 to 7) are shown in Table 2. After the algorithm terminates, i.e. after the 8th iteration of the for-loop (statements 3 to 7), *Cf* gives the closed frequent intervals - viz. [2,4], [2,5], [2,6], [4,5], [4,6], [4,7], [4,8] in the input transaction database along with their respective support values. It follows from Section 4.3 that the set of closed frequent intervals generated above is both correct and complete with respect to the given transaction database (Table 1) for min_sup = 4.

Table 1
Input Transaction Database, min_sup = 4

TID	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*
I	[1,8]	[1,7]	[2,8]	[2,6]	[2,5]	[2,4]	[4,9]	[4,9]	[4,9]	[4,7]	[4,5]

Table 2
Tracing the Closed Frequent Interval Generation Algorithm

Iteration No.	*Value of k*	*Output of of Maxfreq(k)*	*Intervals added to Cf*	*Cf*
1	11	-	-	NIL
2	10	[4,5]	[4,5]	([4,5],10)
3	9	[4,5]	-	([4,5],10)
4	8	[4,6]	[4,6]	([4,5],10),([4,6],8)
5	7	[4,7]	[4,7]	([4,5],10), ([4,6],8), ([4,7],7)
6	6	[2,4], [4,7]	[2,4]	([4,5],10), ([4,6],8), ([4,7],7), ([2,4],6)
7	5	[2,5], [4,8]	[2,5], [4,8]	([4,5],10), ([4,6],8), ([4,7],7), ([2,4],6), ([2,5],5),([4,8],5)
8	4	[2,6], [4,8]	[2,6]	([4,5],10), ([4,6],8), ([4,7],7), ([2,4],6), ([2,5],5),([4,8],5), ([2,6],4)

5. CONCLUSION AND FUTURE WORKS

The notion of closed frequent intervals was defined in a database of intervals. A novel pioneering algorithm was proposed to discover the closed frequent intervals in a database of intervals for a given minimum support threshold. The correctness of the proposed algorithm has been established mathematically. Future works include generalizing the proposed algorithm to discover closed multi-dimensional intervals also.

REFERENCES

[1] Ale J.M., and Rossi G.H.; "An Approach to Discovering Temporal Association Rules", *In Proc. of 2000 ACM Symposium on Applied Computing (2000).*

[2] Ozden B., Ramaswamy S. and Silberschatz A., "Cyclic Association Rules". *Proc. of the 14th Int'l Conf. on Data Engineering*, USA, pp. 412-421 (1998).

[3] Zimbrado G., Moreira de Souza J., Teixeira de Almeida, V. and Arauja de Silva, W.; "An Algorithm to Discover Calendar-based Temporal Association Rules with Item's Lifespan Restriction", *In Proc. of the 8th ACM SIGKDD*, (2002).

[4] Agrawal R., Ramakrishnan, S., "Fast Algorithms for Mining Association Rules in Large Databases", *Proceedings of the 20th International Conference on Very Large Data Bases*, pp. 487-499, September 12-15, (1994).

[5] Agrawal R., Ramakrishnan S.; "Mining Sequential Patterns", *Proceedings of the Eleventh International Conference on Data Engineering*, pp. 3-14, March 06-10, (1995).

[6] Lin J., "Mining Maximal Frequent Intervals". *In: Proceedings of 2003 ACM Symposium on Applied Computing*, pp. 426-431. ACM, New York (2003).

[7] Dutta M., Mahanta A.K.; "An Efficient Method for Construction of I-tree". *In: Proceedings of National Workshop on Design and Analysis of Algorithm (NWDA)* (2010).

[8] Garofalakis M., Rastogi R., Shim K.; "Spirit; Sequential Pattern Mining with Regular Expression Constraints". *In Proc. of the 1999 Intl Conf.* Very Large Databases (VLDB'99), pp. 223-234 (1999).

[9] Masseglia F., Cathala F., Poncelet P.; "The psp Approach for Mining Sequential Patterns". *In Proc. 1998 European Symp. Principle of Data Mining and Knowledge Discovery* (PKDD'98), pp. 176-184 (1998).

[10] Bohlen M.H., Busatto R., Jensen C.S.; "Point Versus Interval-based Temporal Intervals". *In Proc. of the Fourteenth International Conference in Data Engineering*, pp. 192-200 (1998).

[11] Hoppner F., Klawonn F., "Finding Informative Rules in Interval Sequences". *In Proc. of the 4th International Conference on Advances in Intelligent Data Analysis*. LNCS, **Vol. 2189**, pp. 125-134 (2001).

[12] Wu S., Chen Y., "Mining Non-ambiguous Temporal Patterns for Interval-Based Events". *IEEE Transactions on Knowledge and Data Engineering*, **19(6)**, pp. 742-758 (2007).

[13] Kam P., Fu A.W., "Discovering Temporal Patterns for Interval-based Events", *In Proc. of the Second International Conference on Data Warehousing and Knowledge Discovery.* pp. 317-326 (2000).

[14] Mazarbhuiya F.A., Khaleel M.A., Mahanta A.K., Baruah H.K., "Mining Maximal Dense Intervals from Temporal Interval Data". *International Journal of Computer Science and Information Security*, pp. 102-107 (2011).

[15] Khaled M. Elbassioni. "Finding All Minimal Infrequent Multi-dimensional Intervals". *In LATIN*, pp. 423-434 (2006).

[16] Pei J., Han J., Mao, R., "Closet: An Efficient Algorithm for Mining Frequent Closed Itemsets". *In SIGMOD Int'l Workshop on Data Mining and Knowledge Discovery*, May 2000.

[17] M.J. Zaki and C.-J. Hsiao. "ChARM: An Efficient Algorithm for Closed Association Rule Mining". *Technical Report 99-10, Computer Science Dept., Rensselaer Polytechnic Institute*, October 1999.