Mining Rare Sequential Patterns in Large Transaction Databases

Weimin Ouyang
Department of Computer Teaching
Shanghai University of Political Science and Law
Shanghai, China
oywm@shupl.edu.cn

***Abstract*—Sequential pattern is an important research topic in data mining and knowledge discovery. During the discovery of sequential patterns, only the frequent sequences are considered while all the infrequent sequences are ignored. However, some infrequent patterns can provide very useful insight view into the data set and a new kind of knowledge discovery problems called as rare sequential pattern and its discovery algorithm are proposed in this paper. Experiments on the synthetic data set show that the proposed algorithm is efficient and scalable.**

Keywords—data mining, algorithm, rare sequential pattern, sequence

I. INTRODUCTION

Like association rules [1], sequential pattern [2] is an important research topic in data mining and knowledge discovery, which is firstly proposed by R.Agrawal. While association rules mining is to find the intra-transaction relations, sequential patterns mining is to find the inter-transactions relations. A sequential pattern is formed as (A, B), where A and B are disjoint item sets, and its support is no less than a user-specified minimum support. The sequential patter (A, B) means that if A is in a transaction, then B would be in another transaction with high probability.

Most of the researches on mining sequential patterns are confined to the frequent sequences, whose support is no less than the predefined minimum support threshold. In some situations, however, it may be interesting to search for "rare" itemsets, i.e. itemsets that do not occur frequent in the data (contracting frequent itemsets) [3], and a new kind of knowledge discovery problems called as rare association rules has been proposed [4,5,6,7]. Inspired by the idea of mining rare association rules, Yu lei, et al, put forward the problem of mining rare sequential patterns [8]. However, they wrongly introduced the concept of window based on the number of items, which made its algorithm have no real value. As far as I know, except literature [8], there are no other researchers involved in this problem. In order to settle the problem of mining rare sequential patterns, we put forward rare sequential pattern and its discovery algorithm in this paper.

The paper is organized as follows. The definitions for rare sequential pattern are given in Section 2. In Section 3, we describe the discovery algorithm for mining rare sequential patterns. Section 4 presents our primary experimental results. The conclusion and future works are made in the last section.

II. RELATED WORKS

Recently, mining rare association rules has attracted numerous researches. There are two different types of rare association rules mining approaches: level-wise and tree based. Current rare itemsets mining approaches which are based on level-wise exploration of the search space are similar to the Apriori algorithm. MS-Apriori [9], Apriori-Inverse [3], Rarity [10], ARIMA [5] and AfRIM [6] are designed to discover rare itemsets, they all use level-wise algorithm similar to Apriori. Tsang et al. [7] proposed a RP-Tree algorithm to handle these issues. RP-Tree avoids the expensive itemset generation and pruning steps by using a tree data structure to find rare patterns.

Sequential pattern mining in transaction dadabase is firstly proposed by R.Agrawal et al, which has been well studied in. Apriori [2] and GSP [11] are the famous algorithms for mining sequential patterns based on association rules mining technique. SPADE [12], which was proposed by Zaki, systematically studied the problem of sequence lattice in sequential pattern mining. Prefix [13], proposed by Pei and Han et al, is very efficient, which does not need to generate candidate patterns and identify their occurrences but grows patterns as long as the current item is frequent in the projected dataset. Yu lei, et al, proposed the problem of mining rare sequential patterns [8]. However, they wrongly introduced the concept of window based on the number of items. As far as I know, except literature [8], there are no other researchers involved in this problem.

III. PROBLEM DEFINITIONS

A. Sequential Patterns

Let $I = \{i_1, i_2, \ldots, i_m\}$ be a set of literals called items. Let the database $D = \{ t_1, t_2, \ldots, t_n \}$ be a set of n transactions, where each transaction is a subset of I. A non-empty subset of I is called itemset. An itemset containing k items is called k-itemset. The support of an itemset X denoted as sup(X) is defines as the fraction of all transactions containing X in D. An itemset is frequent if its support is greater than a user-specified threshold minsupp.

A sequence is an ordered list of itemsets such as $s = (s_1, s_2, \ldots, s_u)$, where each itemset s_i is an element of the sequence. A sequence is said to be non-empty if it contains at least one element. An item can appear only once in an element, but can

occur multiple times in different elements of a sequence. Items in an element are assumed to be sorted in lexicographic order. A sequence with k items, where $k = \sum_j |s_j|$, is called a k-sequence, where $|s_j|$ denotes the number of items in itemset s_j.

A sequence $t = (t_1, t_2, \ldots, t_v)$ is called a sub-sequence of s if there exist integer $1 \le j_1 < j_2 < \ldots < j_v \le u$ such that $t_1 \subseteq s_{j1}$, $t_2 \subseteq s_{j2}$, … , $t_v \subseteq s_{jv}$. A sequence database SD is a set of tuple <sid,t> where sid is the sequence identifier and t is a sequence. A sequence database SD is a set of tuple <sid,t> where sid is the sequence identifier and t is a sequence. A tuple <sid,t> is said to contain a sequence s if s is a sub-sequence of t. The support of a sequence s, sup(s), is defined as the fraction of all tuples in SD that contain s. A sequence is called sequential pattern if the support of the sequence is no less than the predefined minimum support threshold.

The algorithm proposed by Agrawal and Srikant to discover sequential patters from large transaction databases is divided into five phases [1] . (1) Sort phase. The transaction database is sorted by customer ID as the major key and transaction time as the minor key. The phase converts the original transaction database into a database of customer sequences; (2) Large itemset phase. The set of all the large itemsets are found from the customer sequences database by the similar method with the process of mining association rules. Note that the counting of itemset is for customer sequence not for transaction. When an itemset occurs more than one time in a customer sequence, it is counted just once for this customer sequence. (3) Transformation phase. In this phase, each large itemset is mapped to a contiguous integer and the original customer sequences are transformed into the mapped integer sequences. (4) Sequence phase. The set of transformed integer sequences are used to find frequent sequences among them. (5) Maximum phase. The maximally large sequences are derived and output to users.

B. Rare Sequential Patterns

Inspired by the idea of rare association rules, we put forward the concept of rare sequential patterns. The rare sequential pattern is defined as follows:

Definition 1: Given a user-defined maximum support threshold maxs, a user-defined minimum support threshold mins, A sequence X is called rare sequential pattern, if and only if $sup(X) < maxs \cdot w$, $sup(X) \ge mins \cdot w$.

IV. DISCOVERY ALGORITHM FOR MINING RARE SEQUENTIAL PATTERNS

According to the definitions of rare association rules in last section, we propose an algorithm to discover rare association rules in data stream called MWRAR-SW (Mining weighted rare Association Rules in a Sliding Window). In the proposed algorithm, for each item X in the current sliding window SW, we construct a bit-sequence with w bits denoted as Bit(X). If an item X is in i-th transaction of the current window SW, i-th bit of Bit(X) is set to be 1; otherwise, it is set to be 0. The process is called bit-sequence transform.

Sequential pattern is an important research topic in data mining and knowledge discovery. Traditional algorithms for mining sequential patterns are built on the discovery of frequent sequences, and only frequent sequential patterns are discovered. Recently, researchers have recognized that some infrequent patterns can provide very useful insight view into the data set and a new kind of knowledge discovery problems called as rare association rules has been proposed. Similarly, we put forward a new kind of discovery problem called as rare sequential patterns and a discovery algorithm for mining rare sequential patterns. Notations used in this algorithm are described as Table 1.

TABLE 1: NOTATIONS

Notation	meaning
n	The total number of transactions in database
m	The total number of items
c	The total number of customers
C_k	the set of candidate itemsets with k items
R_k	the set of rare itemsets with k items
SC_k	the set of candidate sequences with k itemsets
SR_k	the set of frequent sequences with k itemsets
maxs	The predefined maximum support threshold
mins	The predefined minimum support threshold

The proposed mining algorithm is described as follows.

Algorithm: MRSP (Mining rare sequential patterns)

Input: A body of n transaction data, each consists of customer ID, transaction time and the purchased items, a predefined maximum support threshold maxs and a predefined minimum support threshold mins;

Output: a set of rare sequential patterns RSP;

(1) The transaction database is sorted by customer ID as the major key and transaction time as the minor key.
(2) $L = \varnothing$;
(3) Scan the database D for finding rare 1-itemset R_1, where $R_1 = \{ x \mid x \in I, sup(x) < maxs \wedge sup(x) \ge mins \}$.
(4) $R = R \cup R_1$;
(5) For ($k=2; R_{k-1} \ne \varnothing$; k++) {
(6) C_k = Candiate_Gen(R_{k-1});
(7) For each transaction sequence t in D do {
(8) for each candidate k-itemset $c \in C_k$ do
(9) if c is contained in transaction sequence t
(10) c.count++;
(11) };
(12) $R_k = \{c \mid c \in C_k, sup(c) < maxs \wedge sup(c) \ge mins\}$;
(13) $R = R \cup R_k$;
(14) }
(15) Map each rare itemset $\in$ L to a contiguous integer and put it in the rare 1-sequence set SR_1.
(16) Transform each customer sequence using the integer representation.
(17) RSP= $\varnothing$;
(18) For ($k=2; SR_{k-1} \ne \varnothing$; k++) {
(19) SC_k = Candiate_Gen(SR_{k-1});
(20) For each transaction sequence t in D do {
(21) $Temp_t$ = k-itemsets in both t and SC_k;

(22) For each itemset i in $Temp_t$ do i.count++;
(23) };
(24) $SR_k = \{c | c \in SC_k, sup(c) < maxs \wedge sup(c) \geq mins\}$;
(25) $RSP = RSP \cup SR_k$;
(26) }
(27) Transform each k-sequence in RSP , k> 2, into sequences of original items and output them to users as indirect sequential patterns.

Step 1 is sorting phase to convert the original transaction database into a database of customer sequences.

Step 2~14 is rare itemset phase. The set of all the rare itemsets are found from the customer sequences database by the similar method with the process of mining sequential patterns. The generation of candidate large k-itemsets is carried by function Candiate_Gen(R_{k-1}) in the same way as in the Apriori algorithm. Compared with mining algorithm for association rules such as Apriori, an important difference is that counting of itemset is for customer sequence not for transaction. When an itemset occurs more than one time in a customer sequence, it is counted just once for this customer sequence.

Step 15~16 is transformation phase. In this phase, each rare itemset is mapped to a contiguous integer and the original customer sequences are transformed into the mapped integer sequences.

Step 18~26 is a loop. It generates SR_k for each k>=2, where SR_k are the all rare k- sequences after taking kth pass over the transformed customer sequence database D. The loop termination condition is SR_{k-1}.

The passing over the transformed customer sequence database D is composed of two parts. One is generation of candidate k-sequences, the other is support counting for each k-sequence. Candidate k-sequence is generated by linking two of each rare (k-1)-sequence in SR_{k-1}. SC_k is composed of all of the Candidate k-sequence. Then take a pruning step by deleting all k-sequences $c \in SC_k$ such that some (k-1)-sequence of c is not in SR_{k-1}.

SR_k is the set of all the frequent k-sequences that satisfy the maximum support threshold and the minimum support threshold.

Step 27 is to transform each element in RSP into corresponding sequence of original items and output them to users as rare sequential patterns.

V. EXPERIMENT

In this section, we evaluate the performance of our proposed algorithm for mining indirect temporal sequential patterns. The computation environments are i5-3470, 4G RAM, Windows 7 operating system. The algorithm is implemented with C++. The synthetic experiment data set is generated by Assocgen [2] program of IBM Almaden research center. The meanings of used parameters are showed in Table 2.

We set parameters C=10, T=5, S=4, I=2.5, NS =500, NI =2500, N =10000, total number of customers D=100000, and the generated dataset is named as C10T5S4I25. Figure 1 shows the algorithm executing time variance with the maximum support maxs and the minimum support mins decreasing from 1% to 0.2% and from 0.5% to 0.1%,respectively, where the maximum support maxs and the minimum support mins have been satisfied as maxs = 2*mins. It demonstrates that the algorithm increases with the declining of maxs and mins.

To examine the scalability of algorithm we increased the numbers of customer D from 50,000 to 150000, with maxs = 1%. The results are shown in Figure 2. The executing time is increased almost linearly with the increasing of dataset size. It can be concluded our algorithm has a good scalable performance.

TABLE 2: PARAMETERS

Symbol	meaning
D	Number of customers(=size of database)
C	Average number of transactions per Customer
T	Average number of items per Transaction
S	Average length of maximal potentially large Sequences
I	Average size of Items in maximal potentially large sequences
N_S	Number of maximal potentially large Sequences
N_I	Number of maximal potentially large Itemsets
N	Number of items
D	Number of customers(=size of database)

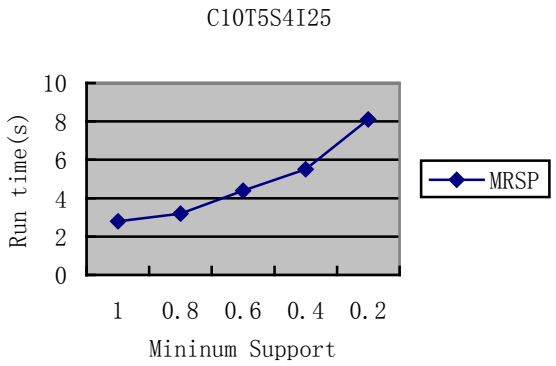

Figure 1: Running time.

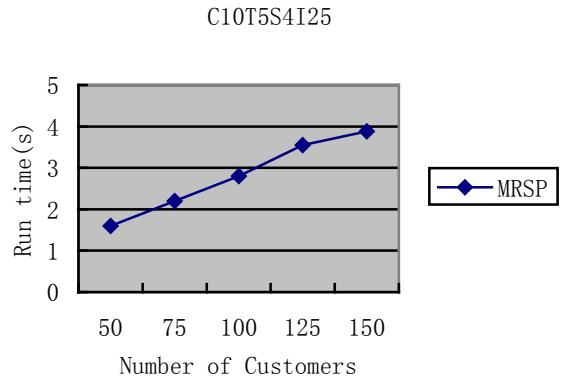

Figure 2: Scale-up: Number of customers.

VI. CONCLUSIONS AND FUTURE WORKS

In this paper, we addressed the one of limitations of traditional sequential patterns mining, which is only frequent sequence to be considered. Inspired of the idea of rare association rules, we put forward an another new kind of discovery problem called as rare sequential patterns, and implemented the discovery algorithm for mining rare sequential patterns with C++ in Windows XP environment. The primary experiments demonstrated that the algorithm is efficient and scalable. As the future works, we plan to address rare sequential patterns from transaction databases with quantitative values and items with concept hierarchical relationships.

REFERENCES

[1] Agrawal R,Srikant R. Fast Algorithms for Mining Association Rules In the Proc. of the 20th International Conference on VLDB. Santiago, 1994. pp.487~499.

[2] Agrawal R,Srikant R. Mining sequential patterns. In: Proceedings of International Conference on Data Engineering, Taibei, Taiwan, pp3-14, March 1995.

[3] Koh, Y.S., Rountree, N.: Finding Sporadic Rules Using Apriori-Inverse. In: Ho, T.-B., Cheung, D., Liu, H. (eds.) PAKDD 2005. LNCS (LNAI), vol. 3518, pp. 97-106. Springer, Heidelberg (2005).

[4] Szathmary, L., Napoli, A., Valtchev, P.: Towards rare itemset mining. In: Proceedings of the 19th IEEE International Conference on Tools with Artificial Intelligence, ICTAI 2007, vol. 1, pp. 305‑312. IEEE Computer Society, Washington, DC (2007).

[5] Torino L, Sibelius G, Barolo C, "A fast algorithm for mining rare itemsets". In the Proc. of the Ninth International Conference on Intelligent Systems Design and Applications, pp.1149-1155, 2009.

[6] Adda M, Wu L, Feng Y, "Rare itemset mining", In the Proc. of the Sixth International Conference on Machine Learning and Applications, pp.73-80, 2007.

[7] Tsang S, Koh Y.S, Dobbie G, "RP-Tree: Rare Pattern Tree Mining", In the Proc. of DaWaK 2011, vol. 6862, pp. 277-288, 2011.

[8] Yu Lei, Man Li, Weisong Hu, Guojie Song, Kunqing Xie, "Efficient methods for rare sequential pattern mining", Journal of Frontiers of Computer Science and Technology, 9(4), pp.429-437,2015

[9] Liu B, Hsu W, Ma Y, "Mining association rules with multiple minimum supports", In: Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 337‑341, 1999.

[10] Szathmary L, Napoli A, Valtchev P, "Towards rare itemset mining", In the Proc. of the 19th IEEE International Conference on Tools with Artificial Intelligence, vol. 1, pp. 305-312, 2007.

[11] R.Srikant, R.Agrawal, "Mining sequential patterns: generalizations and performance improvements", Proc. 15th Int'l Conf. Extending Database Technology(EDBT'96), pp.3-14, Mar. 1996.

[12] M.J.Zaki, "Spade: An efficient algorithm for mining frequent sequences", Math. Learn. 42(1-2), pp.31-60,2001.

[13] J.Pei, J.Han, et al., "Mining sequential patterns by pattern-growth: The prefixspan approach", IEEE Trans. Knowledge Data Eng., 16(11), pp.1424-1440, 2004.