Overview of Redundancy Free Association Rule Mining

A K Chandanan
Department of Computer Science and IT
Jayoti Vidyapeeth Women's University, Jaipur (India)

M K Shukla
Department of Computer Science & Engineering,
Sunder Deep Group of Institution, Ghaziabad (India)

ABSTRACT
Association rule mining is a way to find relations or co-relations among a set of information available. The aim to generate rules for giving multiple data from various databases. Analysis of data can be possible with the help of sequential access of data from database. In case of sequential access of data it may cause multiple times same rules to be generated. It is desired to find a solution to get out of those unnecessary association rules due to the complex characteristics of serial data. Although many numbers of serial association rule with the use of either sequence or temporal constraint as prediction model, these two models did not consider with the repetition during the process of rule mining for the database. The goal of this paper is to propose a method for redundancy free serial association rule mining.

Keywords
Association rule mining, Sequence Creator, Redundancy free serial rule mining.

1. INTRODUCTION
Association rule discovery[13], a successful and important mining task, aim at uncovering all frequent patterns among transactions composed of data attributes or items[16]. Results are presented in the form of rules between different sets of items, along with metrics like the joint and conditional probabilities of the antecedent and consequent, to judge a rule's importance [18]. A closed set [17] contains its own boundary. In other words, if you are "outside" a closed set, you may move a small amount in any direction and still stay outside the set. Note that this is also true if the boundary is the empty set, e.g. in the metric space of rational numbers, for the set of numbers of which the square is less than two. Any intersection of closed sets is closed and any union of finite many closed sets is closed. In particular, the empty set and the whole space are closed. In fact, given a set Y and a collection A of subsets of Y that has these properties, and then A will be the collection of closed sets for a unique topology on Y. The intersection property also allows one to define the closure of a set B in a space Y, which is defined as the smallest closed subset of Y that is a superset of B. Specifically; the closure of B can be constructed at the intersection of all of these closed supersets. Sequential pattern mining, which is the process of extracting certain sequential patterns whose support exceeds a predefined minimal support threshold, has been studied widely in the last decade in the data mining community However, and less work, has been done on sequential association rule mining[1][7].

Only in recent years, several prediction models which introduced the concept of sequential association rule mining have been proposed [1], most of which use sequence and temporal constraints in generating association rules. In the classical association rule mining[6], the resulting rule set can easily contain thousands of rules in which many of the rules are redundant and are useless in practice. While in the case of sequential association rule mining, things get even worse This is because the same set of items with different ordering yields different sequential patterns in sequential pattern mining which makes the number of frequent sequential patterns usually much larger than the number of frequent item sets generated from a dataset of a similar size. When rules are in rapid growth for the set of association rules , especially as we lower the frequency requirements. The larger frequent item sets causes for more the number of rules to be generated for the dataset, many of which are redundant.

In many applications such as web log data, system traces, purchase histories, financial market data typically is represented in the form of sequences [2]. Sequence data can be a consequence of employing a natural temporal ordering among. An important research in the field of data mining, recently mining sequential data has drawn more and more attention to researchers in the data mining field. In the last decade, many algorithms and techniques have been proposed to deal with the problem of sequential pattern mining including. A priority-based approaches such as GSP and SPADE and pattern-growth based approaches [15]. These existing approaches mainly discuss how to efficiently generate sequential patterns, and do not pay much attention to the quality of the discovered patterns, in particular, all of these approaches suffer from the problem that the volume of the discovered patterns and association rules could be exceedingly large, but many of the patterns and rules are actually redundant and thus need to be pruned.

2. RELATED WORK
Recently, some researchers proposed definition of association rule mining it is a way to find interesting associations among large sets of data items have been reviewed [14]. Association rule mining [6], which aims to extract interesting correlations and associations among sets of items in large datasets, has two phases: extracting frequent item sets and generating association rules from the frequent item sets with the constraints of minimal support and minimal confidence [8]. The rules with a confidence value larger than a user-specified minimum confidence threshold is considered interesting or useful. There are two basic measures used in association rule mining, support and confidence. Support is a measure that defines the percentage of records in the dataset that contain $A \cup B$ to the total number of records. Confidence is a measure that defines the percentage of records in the dataset that contain $A \cup B$ to the total number of record that contain just A. The confidence value serves as a measure of the

strength or precision of the rule. Apriori is the first association rule mining algorithm that pioneered the use of support-based pruning to systematically control the exponential growth of candidate item sets .In computer science and data mining; Apriori is a classic algorithm for learning association rules. Apriori is designed to operate on databases containing transactions [20]. Other algorithms are designed for finding association rules in data having no transactions, or having no timestamps. As is common in association rule mining, given a set of item sets, the algorithm attempts to find subsets which are common to at least a minimum number C of the item sets. Apriori uses a "bottom up" approach, where frequent subsets are extended one item at a time and groups of candidates are tested against the data. The algorithm terminates when no further successful extensions are found. Apriori uses breadth-first search and a tree structure to count candidate item sets efficiently. It generates candidate item sets of length L from item sets of length L − 1. Then it prunes the candidates which have an infrequent sub pattern. According to the downward closure lemma, the candidate set contains all frequent L-length item sets. After that, it scans the transaction database to determine frequent item sets among the candidates [19][15].

2.1 Support-confidence

The most common used method in association rule mining to determine interestingness from the dataset is a generality and reliability based on two measures known as the support and confidence which can lead to interesting rules being found by setting low support thresholds [11]. In a support and confidence approach, the support measures or defines the range of the rule and the confidence measures the precision or accuracy of the rule .Support was chosen as it represents statistical significance [5] and users are usually only interested in the rules that have a support/significance above a certain threshold. The calculation of support assumes statistical independence and the support - confidence approach is targeted at finding qualitative rules.

Example: suppose a big shopping mall sales data by stock maintenance unit for each item, and thus is able to know what items are typically purchased together. Apriori [20] is a moderately efficient way to build a list of frequent purchased item pairs from this data. Let the database of transactions consist of the sets {1,2,3,4}, {1,2}, {2,3,4}, {2,3}, {1,2,4}, {3,4}, and {2,4}. Each number corresponds to a product such as "butter" or "bread". The first step of Apriori is to count up the frequencies, called the supports, on each member item separately: This table explains the working of Apriori algorithm:

Item	Support
1	3
2	6
3	4
4	5

In Apriori, the minimum support level for qualifying a pattern as "frequent" might be defined, which depends on the context. In this example, let min support = 3. Therefore, all datasets are frequent. The next step is to generate a list of all 2-pairs of the frequent items. Had any of the above items not been frequent, they wouldn't have been included as a possible member of possible 2-item pairs. In this way, Apriori prunes the tree of all possible sets. In the next step we have again selected only these items which are frequent:

Item	Support
{1,2}	3
{2,3}	3
{2,4}	4
{3,4}	3

and generate a list of all 3-triples of the frequent items. In the example, there are no frequent 3-triples. Most common 3-triples are {1, 2, 4} and {2, 3, 4}, but their support is equal to 2 which is smaller than our min support. Since all of the Apriori-based mining algorithms have time or space costing problems when handling a huge number of candidate sets and a large database [12], a new method which avoids candidate generation-and-test and utilizes a new data structure to reduce cost . It is FP-Tree algorithm. [3] The basic idea of the FP-growth algorithm can be described as a recursive elimination scheme: in a preprocessing step delete all items from the transactions that are not frequent individually, i.e., do not appear in a user-specified minimum number of transactions. The pattern - growth approach is more efficient and scalable than other approaches, such as Apriori3, and is effective in the mining dense database. Sequential pattern mining5 is the process of extracting certain sequential patterns whose support exceeds a predefined minimal support threshold. Since the number of sequences can be very large, and users have different interests and requirements, to get the most interesting sequential patterns usually a minimum support is predefined by the users. By using the minimum support we can prune out those sequential patterns of no interest, consequently making the mining process more efficient. Introduced another metric called surprise to measure the interestingness of sequences

3. METHEDOLGY

In this paper we give the algorithm of non redundant association rule mining. From sequential patterns, we can extract relations between the sets of sequences in the format of sequential association rules. However, the huge number of sequential rules will cause several problems. The first issue is the large quantity of low quality rules that are almost meaningless and will not give people any useful information. The second issue is that the rule generation cost of the full sequential rules is also quite high, even for a sparse dataset. Moreover, sometimes it is even impossible to use for mining full sequential rules for dense datasets. So it is quite reasonable to seek a concise and non-redundant representation of sequential rules.

Redundancy free association rule mining: A redundancy free sequential mining method is adapted to extract valuable information from the user navigation sessions generated by web usage mining technique. In non-sequential association rule miningfield6 defines a condensed representation of association rules; the representation was characterized by frequent closed item sets and their generators. And the research aimed to present rules with minimal preceding and maximal successive. This technique can remove a significant amount of redundant association rules [9] to help improve the quality of the mining result. We extend the redundancy free association rule theory into the sequential mining area. We

first give the definitions of the redundant sequential rules, and then we introduce the smallest-largest sequential basis for redundancy free rule representation which requires both sequential generators and closed sequential patterns. Finally we provide an algorithmic approach that might be efficiently extract redundancy free sequential rules.

Definition [4] - Redundant Sequential Association Rules: Suppose A → B and A` →B` be two sequential rules with confidence values C and C`, respectively. A →B is said a redundant rule to A` →B` if A` belong to A; B belongs to B`, and C ≤ C'.

Definition - Confidence of sequential rules: Suppose A, B are two different sequences in a transaction database. The confidence of rule A U B is defined as support (A, B) /support (A). Notify the support (A, B) represents the number of sequences that contain both A and B in a transaction database. Sequential smallest-largest rules and Sequential rules share the same attribute as non-sequential rules, such as transitivity, support, and confidence. The main distinction is that sequential associations describe sequential relationships between items with the redundant rule definition

Definition [10]- Smallest-Largest sequential association rules: Suppose R be the set of sequential association rules. A sequential rule r: $s_1 \rightarrow s_2 \in R$ is a Smallest-Largest sequential rule if not $\exists \rho \rightarrow$: $s_1` \rightarrow s_2` \in R$ with support(s`) = support(s); confidence(s`) = confidence(s), $s \subseteq s_1$, and $s_2` \subseteq s_2`$.The Smallest-Largest sequential association rules are the redundancy free sequential rules having minimum preceding and largest successive. r is a Smallest-Largest sequential rule if no other sequential rule r` has the same support and confidence, and it has an preceding that is the subsequence of the preceding of s and a successive that is a super sequence of the consequent of r. In this paper there are two new bases are defined that is suitable for sequential data. These two are sequential Smallest-Largest exact base and sequential Smallest-Largest approximate base. These two bases also formed a concise representation of sequential association rules. Suppose that x, y, z are sequences, $x \subset y \subset z$, closure (x) = closure (y) = closures (z), then the two rules a_1: x→ (z\x) and a_2: y→ (z\x), where z\y denotes a subsequence of z by removing the sequence y, will have the same confidence, the antecedent of a_1 is shorter than that of a_2 and the consequent of a_1 is longer than that of a_2. If z is a closed sequence and x is a generator, i.e., z=closure (z) and x is the minimum sequence which has the same closure as z, $x \subset (z\backslash x)$ will have the shortest preceding and longest successive among the rules $y \subset (z\backslash y)$ where $x \subset y \subset z$. Therefore, similar to Smallest-Largest exact rules which are generated using a closed item set and its generator, the sequential exact rules can be generated using a closed sequence and its sequential generators Since the closure (x) =closure (z), the confidence of $x \subset (z\backslash x)$ is 1. The sequential Smallest-Largest exact rules are defined as follows:

Definition - Sequential Smallest-Largest Exact Base: Suppose, Closed be the set of closed sequential patterns, and for each closed sequential pattern CP, let Gen_CP be the set of sequential generators of CP. The sequential Smallest-Largest exact base is:

Sequential Smallest-Largest Exact = {r: x → (CP\x) | CP ∈ Closed ∧ x ∈ Gen_CP ∧ x ≠ CP}

Definition - Sequential Smallest-Largest Approximate Base: Suppose, Closed be the set of closed sequential patterns, Gen_all be the set of all sequential generators of closed sequential patterns in Closed. The sequential Smallest-Largest approximate base is:

Sequential Smallest-Largest Approx = {r: x → (CP\x) | CP ∈ Closed ∧ x ∈ Gen_all ∧ Closure (x) ≠ CP}

Algorithm: RF Rule Mining (C, mn_sup, mn_confidence)
Input: mn_sup, mn-confidence, Closed sequence set C
Output: Rule set R (redundancy free (RF) sequential rules)

i. Suppose C` = closed sequence set of C, GN = generator set of C
ii. For each sequence pattern x ∈ Γ
iii. For each closed sequence pattern CP ∈ Λ
iv. d: = CP/x
v. r: x →α, support(r) = support (CP), and
Confidence (r) = support (CP) /support (x)
vi. If (confidence(r) >= mn_confidence)
vii. Add rule r into rule set R

For example, if a sequence WX is on a generator, while
The sequence WXYZ is a closed sequence, and then WX U YZ is considered as a redundancy free sequential rule.

4. CONCLUSION

We presented the definition of the redundancy of sequential association rule and proposed the Sequential Smallest-Largest base for the concise representation of redundancy free sequential association rules. According to this basis, we introduced a method for mining redundancy free sequential rules based on sequential creator and closed sequential patterns. By using this method the generated redundancy free sequential rules have the minimum preceding and the maximum successive. In future work, we will explore a sequential pruning mechanism in which only sub rules are used that is confident and that's where not already pruned earlier.

5. REFERENCES

[1] Agrawal R. and Srikant R., Mining sequential patterns, Proceedings of the Eleventh International Conference on Data Engineering 1995 (1995).

[2] Brin S., Motwani R., Ullman J.D., and Tsur S., Dynamic item set counting and implication rules for market basket data, In SIGMOD 1997, Proceedings ACM SIGMOD International Conference on Management of Data, May 13-15, 1997, 255-264 (1997)

[3] Han J., Pei J, Mining Frequent Patterns by Pattern-Growth: Methodology and Implications., ACM SIGKDD (2000)

[4] Ganter B. and Wille R., Formal Concept Analysis: Mathematical Foundations, Springer, Berlin- Heidelberg-New York, 10, (1999) Brown, L. D., Hua, H., and Gao, C. 2003. A widget framework for augmented interaction in SCAPE

[5] Zhao Q., Bhowmick, S. S., Association Rule Mining: A Survey. Nanyang Technological University, Singapore.(2003)

[6] Kotsiantis S,Kanellopoulos D,Association Rules Mining: A Recent Overview ,GESTS International Transactions

on Computer Science and Engineering, Vol.32 (1), 2006, pp. 71-82

[7] Gaul W. and Schmidt-Thieme L., Mining Generalized Association Rules for Sequential and Path Data, Proceedings of the 2001 IEEE International Conference on Data Mining (2001) Spector, A. Z. 1989. Achieving application requirements. In Distributed Systems, S. Mullender

[8] Agrawal R. and Srikant R., Fast algorithms for mining association rules in large databases, Proceedings of 20th International Conference on Very Large Databases (1994)Tavel, P. 2007 Modeling and Simulation Design. AK Peters Ltd.

[9] Ashrafi M.Z., Taniar D. and Smith K., Redundant association rules reduction techniques, International Journal of Business Intelligence and Data Mining (2007)

[10] Guo S., Liang Y., Zhang Z. and Liu W., Association Rule Retrieved from Web Log Based on Rough Set Theory. Proceedings of the Fourth International Conference on Fuzzy Systems and Knowledge Discovery, 24-27 Aug 2007, Vol. 3 ,129 – 135.

[11] Han, J., Kamber, M. , Mining Frequent Patterns, Associations, and Correlations. In D. D. Cerra (Ed.), Data Mining: Concepts and Techniques , 2nd ed., pp. 227-283, 2006, San Francisco, USA: Morgan Kaufmann Publishers

[12] Umarani V, Punithavalli M, A Study on Effective Mining of Association rules from Huge Databases, IJCSR International Journal of Computer Science and Research, 2010, Vol. 1 Issue 1,30-34(2010)

[13] Ceglar A., Roddick J F, Association Mining,. ACM Computing Surveys (C SUR), 38(2).(2006).

[14] Chandanan A K,Shukla M K ,Data mining for qualitative dataset Using association rules: A review, International Journal of Advanced Research in Computer Science and Electronics Engineering Volume 2, Issue 2, February 2013, ISSN: 2277 – 9043,Page 231-238.

[15] Santosh B,Rukmani K, Implementation of Web Usage Mining Using Aproiri and FP Growth Algorithms",Int. J. of Advanced Networking and Applications,Volume:01, Issue:06, Pages: 400-404 (2010)

[16] Tanna P, Ghodasara Y, Foundation for Frequent Pattern Mining Algorithms Implementation ,International Journal of Computer Trends and Technology (IJCTT) – Volume 4 Issue 7 - July 2013

[17] http://en.wikibooks.org/wiki/Data_Mining_Algorithms_In_R/Frequent_Pattern_Mining/The_Apriori_Algorithm

[18] N Raheja,R Kumar,Optimization of Association Rule learning in distributed database using clustering techniques ,International Journal on Computer Science and Engineering (IJCSE),ISSN : 0975-3397Vol. 4 No. 12 Dec 2012 pg 1874-1880

[19] http://en.wikipedia.org/wiki/Apriori_algorithm

[20] http://www.gabormelli.com/RKB/Apriori_Algorithm