An Introduction to Effective Sequential Pattern Mining

[1]Virendra Mishra [2]Abhishek Raghuwansi
[1] Department of Information Technology, MIT, Ujjain (M.P.)
[2] Department of Computer Science & Engineering, MIT, Ujjain (M.P.)

Abstract— Sequential rule mining is a favorite topic of research for many researchers. Sequential rule mining has been applied in several domains such as stock market analysis, weather observation and drought management etc. There are many techniques for the sequential rule mining. But still there is a lot of scope for the improvement in terms of efficiency and scalability. This paper presents a review of the some common and popular techniques for mining sequential rules from a data set.

I. INTRODUCTION

Recent developments in computing and automation technologies have resulted in computerizing business and scientific applications in various areas. Turing the massive amounts of accumulated information into knowledge is attracting researchers in numerous domains as well as databases, machine learning, statistics, and so on. From the views of information researchers, the stress is on discovering meaningful patterns hidden in the massive data sets. Hence, a central issue for knowledge discovery in databases, additionally the main focus of this thesis, is to develop economical and scalable mining algorithms as integrated tools for management systems.

Data mining, that is additionally cited as knowledge discovery in databases, has been recognized because the method of extracting non-trivial, implicit, antecedently unknown, and probably helpful data from knowledge in databases. The information employed in the mining method usually contains massive amounts of knowledge collected by computerised applications. As an example, bar-code readers in retail stores, digital sensors in scientific experiments, and alternative automation tools in engineering typically generate tremendous knowledge into databases in no time. Not to mention the natively computing- centric environments like internet access logs in net applications. These databases therefore work as rich and reliable sources for information generation and verification. Meanwhile, the massive databases give challenges for effective approaches for information discovery.

The discovered information will be utilized in many ways in corresponding applications. For instance, distinctive the oft times appeared sets of things in a very retail info will be used to improve the choice creatingof merchandise placement or commercial. Discovering patterns of client browsing and buying (from either client records or net traversals) could assist the modeling of user behaviors for client retention or customized services. Given the specified databases, whether relational, transactional, spatial, temporal, or transmission ones, we have a tendency to could get helpful info once the information discovery method if acceptable mining techniques square measure used.

II. BACKGROUND

If a collection of data sequences is given, within which every sequence may be a list of transactions ordered by the transaction time, the matter of mining sequential patterns [3] is to get all sequences with a user such minimum support. every transaction contains a collection of things. A sequential pattern is an ordered list (sequence) of itemsets. The itemsets that area unit contained within the sequence area unit referred to as parts of the sequence. For a given database D that consists of client transactions every group action consists of the subsequent fields: customer-ID, transaction-time, and therefore the things purchased within the group action. an item-set may be a non-empty set of things, and a sequence is an order list of item-sets. We are saying a sequence A is contained in another sequence B if there exist integers i1.

$$\text{Support} = \frac{\textit{The number of sequence that contains this sequence}}{\textit{The total number of sequences}}$$

A sequence is an ordered list of elements (transactions). Each element contains a collection of events (items). Each element is attributed to a specific time or location. Length of a sequence, |s|, is given by the number of elements of the sequence.

ID	Sequences
1	{1,2},{3},{6},{7},{5}
2	{1,4},{3},{2},(1,2,5,6}
3	{1},{2},{6},{5},{6,7}
4	{2},{6,7},{1,2},{2,3}

Table: A Sequence Database

Considering a minimum support = 50% and minimum confidence = 50%, we get following sequential rules.

ID	SEQUENTIAL RULE	SUPPORT	CONFIDENCE
1	{1,2,3} => {5}	.5	1.0
2	{1} => {3,5,6}	.5	.66
3	{1,2} => {5,6}	.75	.75
4	{2} => {5,6}	.75	.75
5	{1} => {5,6}	.5	.5
..	..	..	..

Table:B SEQUENTIAL RULES

The goal of sequential patterns is to search out the sequences that have larger than or equal to an explicit user pre-specified support. Sometimes the method of finding sequential patterns consists of the subsequent sections: sorting phase, finding the massive item-set phase, transformation section, sequence section, and greatest phase.

III. RELATED WORK:

As we know, data are changing all the time; especially data on the web are highly dynamic. As time passes by, new datasets are inserted; old datasets are deleted while some

other datasets are refreshed. It is transparent that time stamp is an important attribute of each dataset, also it's aristocratic in the process of data mining and it can give us more accurate and useful information. For example, association rule mining does not take the time stamp in account, the rule may Buy A=⇒Buy B. If we take time stamp in account then we can get more accurate and useful rules such as: Buy A implies Buy B within two days, three days four days or a week and a month, or usually people Buy A everyday in a week. The second kind of rules, business decision can be more accurate and useful prediction and consequently make more sound decisions.

However, one important limitation of the algorithms of Das et al.,[3] and Harms et al. [4] comes from the fact that they are designed for mining rules occurring frequently in sequences. As a consequence, these algorithms are inadequate for discovering rules common to many sequences. We illustrate this with an example. Consider a sequence database where each sequence corresponds to a customer, and each event represents the items bought during a particular day. Suppose that one wishes to mine sequential rules that are common to many customers. The algorithms of Das et al. [3] and Harms et al. [4] are inappropriate since a rule that appears many times in the same sequence could have a high support even if it does not appear in any other sequences. A second example is the application domain of this paper. We have built an intelligent tutoring agent that records a sequence of events for each of its executions. We wish that the tutoring agent discovers sequential rules between events, common to several of its executions, so that the agent can thereafter use the rules for prediction during its following execution.

Sequential rule mining has been applied in many domains like stock exchange analysis (Das & Lin [8], Hsieh, Wu & Yang [13]), weather observation (Hamilton & Karimi, [10]) and drought management (Harms & Tadesse, [11], Deogun & Jiang, [9]).

In order to reduce the number of iterations, the efficient bi-directional sequential pattern mining approach namely Recursive Prefix Suffix Pattern detection, RPSP [7] algorithm is furnished. The RPSP algorithm finds first all Frequent Itemsets (FI) according to the given minimum support and transforms the database such that each transaction is replaced by all the FI it contains and then finds the patterns. Further the pattern detected based on ith projected databases, and builds suffix and prefix databases based on the Apriori properties. Recursive Prefix Suffix Pattern will increase the number of frequent patterns by reducing the minimum support and vice versa. Recursion gets deleted when the detected FI set of prefix or suffix assigned database of parent database is ineffective. All patterns that correlate to a particular ith proposition database of transformed database, that formed into a set, that is dis-joint from all the other sets. The resultant set of frequent patterns is the sum of the all disjoint subsets. The proposed algorithm tested on hypothetical and sequence data and obtained results were found all satisfactory. Hence, RPSP algorithm may be applicable to many real world sequential data sets.

IV. CONCLUSION

We have performed a systematic study on mining of sequential patterns in large databases. In this paper, a comprehensive survey over sequential rule mining has been presented. The working, merits and demerits of each algorithm are discussed.

REFERENCES

[1] Tan, kumar " introduction to data mining".
[2] Arun Pujari " Introduction to data mining"
[3] Han and Kamber, 2000
[4] Das., G., Lin, K.-I., Mannila, H., Renganathan, G., and Smyth, P. Rule Discovery from Time Series. In Proc. 4th Int. Conf. on Knowledge Discovery and Data Mining (New York, USA, August 27-31, 1998), 16-22.
[5] Harms, S. K., Deogun, J. and Tadesse, T. 2002. Discovering Sequential Association Rules with Constraints and Time Lags in Multiple Sequences. In Proc. 13th Int. Symp. on Methodologies for Intelligent Systems (Lyon, France, June 27-29, 2002), pp. .373-376.
[6] Mannila, H., Toivonen and H., Verkano, A.I. Discovery of frequent episodes in event sequences. Data Mining and Knowledge Discovery, 1, 1 (1997), 259-289
[7] Dr P padmaja, P Naga Jyoti, m Bhargava "Recursive Prefix Suffix Pattern Detection Approach for Mining Sequential Patterns" IJCA September 2011
[8] King Ip Lin., Heikki Mannila, Gautam Das, Gopal Renganathan, & Padhraic Smyth, 1998. Rule Discovery from Time Series. Proc. 4th Int. Conf. on Knowledge Discovery and Data Mining.
[9] Liying Jiang & Jiternder S Deogun, 2005. Prediction Mining – An Approach to Mining Association Rules for Prediction. Proceeding of RSFDGrC 2005 Conference, pp.98-108.
[10] Hamilton, H. J. & Karimi, K. 2005. The TIMERS II Algorithm for the Discovery of Causality. Proc. 9th Pacific-Asia Conference on Knowledge Discovery and Data Mining, 744-750.
[11] Harms, S. K., Deogun, J. & Tadesse, T. 2002. Discovering Sequential Association Rules with Constraints and Time Lags in Multiple Sequences. Proc. 13th Int. Symp. on Methodologies for Intelligent Systems, pp. .373-376.
[12] Hegland, M. 2007. The Apriori Algorithm – A Tutorial. Mathematics and Computation. Imaging Science and Information Processing, 11:209-262.

