TEMPORAL DATABASES AND FREQUENT PATTERN MINING TECHNIQUES

N.Pughazendi
Research Scholar,
Department of Computer science and Engineering
Manonmaniam Sundaranar University

Dr.M. Punithavalli
Director, Department of Computer Science,
Dr.SNS Rajalakshmi College of Arts and Science,

ABSTRACT

Data mining is the process of exploring and analyzing data from different perspective, using automatic or semiautomatic techniques to extract knowledge or useful information and discover correlations or meaningful patterns and rules from large databases. One of the most vital characteristic missed by the traditional data mining systems is their capability to record and process time-varying aspects of the real world databases. . Temporal data mining, which mines or discovers knowledge and patterns from temporal databases, is an extension of data mining with capability to include time attribute analysis. The pattern discovery task of temporal data mining discovers all patterns of interest from a large dataset. This paper presents an overview of temporal data mining and focus on pattern discovery using temporal association rules.

KEYWORDS : Association Rules, Pattern Discovery, Temporal Data Mining, Temporal Rules,

1. INTRODUCTION

Data mining is the process of exploring and analyzing data from different perspective, using automatic or semiautomatic techniques to extract knowledge or useful information and discover correlations or meaningful patterns and rules from large databases. Using these patterns it is possible for business enterprises to identify new and unexpected trends, subtle relations in the data and use them to increase revenue and cut cost. The techniques and tools are used in various businesses, scientific and engineering companies. Data mining has been proved to be advantageous in many areas. Examples include (i) Finance (to identify patterns that help be used to decide the result of a future loan application), (ii) Satellite research (to identify potential undetected natural resources or to identify disaster situations like oil slicks), (iii) Health care (to predict outbreaks of infectious diseases) and (iv) Market research (to predict trend in purchase). Various data mining techniques and their applications to different domain are discussed by many researchers (Witten and Frank 2000; Han and Kamber 2001; Hand et al 2001).

One of the most vital characteristic missed by the traditional data mining systems is their capability to record and process time-varying aspects of the real world databases. In simple words, the traditional data mining techniques lack in their ability to analyze variation of data over time and treat them as ordinary data. Examples of temporal datasets include stock market data, manufacturing or production data, maintenance data, web mining and point-of-sale records. Temporal data mining, which mines or discovers knowledge and patterns from temporal databases, is an extension of data mining with capability to include time attribute analysis. Due to the importance and complexity of the time attribute, a lot of different kinds of patterns are of interest.

The temporal database usually includes two time aspects, namely, valid time and transaction time. Valid time denotes the time period during which a fact is true with respect to the real world. Transaction time is the time period during which a fact is stored in the database. These two time aspects allow the distinction of three different forms of temporal databases. They are

(i) A historical database stores data with respect to valid time.
(ii) A rollback database stores data with respect to transaction time.
(iii) A bitemporal database stores data with respect to both valid and transaction time, that is, they store the history of data with respect to valid time and transaction time.

A temporal database supports three major datatypes: temporal data, static data, and snapshot data. Regardless of the data type, the temporal data mining algorithms should be transparent and should treat all data as some form of temporal data (Gadia and Nair 1993). Temporal data mining tasks that works with temporal data types can be grouped into (i) prediction (ii) classification (iii) clustering (iv) search & retrieval and (v) pattern discovery (Laxman and Sastry, 2006).

Prediction is the task of forecasting future values of the time series from past or historical samples. In classification, each temporal object or data is assumed to belong to one of the predefined class or category and the primary objective here is to automatically determine the corresponding category for the given input sequence. Clustering of temporal data involves grouping a collection of

time series based on their similarity. It is the most frequently used technique in temporal data mining, as it can automatically find structures or patterns in large data sets that would be otherwise difficult to summarize (or visualize). Searching for sequences in large databases is another important task in temporal data mining. Sequence search and retrieval techniques play an important role in interactive explorations of large sequential databases. The problem is concerned with efficiently locating subsequences (often referred to as queries) in large archives of sequences (or sometimes in a single long sequence).

The pattern discovery task of temporal data mining discovers all patterns of interest from a large dataset and is the main topic of study in this paper. Several works has been proposed by several researchers and academicians and this study aims to report some of them. The remaining of the paper is organized as follows. The concepts behind association rules and temporal association rules are provided in the next section. Section 3 provides a brief review of the various solutions proposed for temporal association rule mining and Section 4 concludes the study with future research directions.

2. BASIC CONCEPTS

This section introduces the concepts behind association rules, temporal association rules and types of temporal association rules.

2.1. Association rules

The pattern discovering task works with patterns (local structure in the data) and frequent patterns (patterns that occur frequently in s dataset). Much of data mining literature is concerned with formulating useful pattern structures and developing efficient algorithms for discovering frequent patterns. The importance of finding frequent patterns is two fold. The first is that they can be used to discover useful rules and the second is that these discovered rules can then be used to discover some interesting regularities in the data. The main aim here is find all useful pattern structures and use efficient algorithms to identify frequently occurring patterns. From these frequent patterns, useful rules can be generated, which can be used to infer knowledge.

A rule consists of an antecedent (left-hand side proposition) and consequent (right-hand side proposition) and states that when the antecedent is true, then the consequent will also be true. Association rules are most frequently used for capturing such correlations (Agrawal and Srikant, 1994). Each association rule is combined with two constraints namely, support and confidence, which are normally used to select interesting rules from the set of all possible rules. Support is defined as the proportion of transaction in the dataset which contain the itemset and confidence is defined as an estimate of the probability of finding the right-hand side of the rule in transactions under the condition that these transactions also contain the left-hand side (Hipp *et al.*, 2000).

2.2. Temporal association rules

A temporal association rule is defined as the frequency of an itemset over a time period T and is the number of transactions in which it occurs divided by total number of transaction over a time period (Equation 1). In a similar fashion, confidence of a item with another item is the transaction of both items over the period divided by first item of that period (Equation 2).

$$\text{Support(A)} = \frac{\text{Frequency of occurrences of A in specified time interval}}{\text{Total no of Tuples in specified time interval}} \quad (1)$$

$$\text{Confidence}(A \Rightarrow B[T_s, T_e]) = \frac{= \text{Support\_count}(A \cup B) \text{ over Interval}}{\text{Occurrence of A in interval}} \quad (2)$$

where T_s is the valid start time and T_e is the valid end time of the temporal data.

2.3. Discovery of Temporal Association Rules

The problem of association rule mining can be described as the selection of interesting rules from a huge set of extracted rules. The various techniques used for this either are statistical methods (Tan *et al.*, 2000) or are considered as classification task that is used to select the interested association rules (Janetzko *et al.*, 2004). Several types of temporal association rules have been proposed as extensions to the traditional Agrawal *et al.* (1993) method. Examples include episode rules (Mannila and Toivonen, 1996), sequence rules (Agrawal and Srikant, 1995), cyclic association rules (Ozden *et al.*, 1998), calendric rules (Dunham, 2003), inter-transaction rules (Lu *et al.*, 2000). Regardless of the type of association rule, the main aim of the discovery process is to identify important or rules that help in data mining. Roddick and Spiliopoulou (2002) have presented a comprehensive overview of techniques for the mining of temporal data using three dimensions: data type, mining operations and type of timing information (ordering). On the other hand, Antunes and Oliviera (2001) base their classification on representation, similarity and operations. In general, an effective algorithm to discover association rules is the apriori. Several variations that enhance these algorithms for temporal database also exist.

3. TEMPORAL ASSOCIATION RULE MINING

Temporal association rule mining can be performed in four sequential steps, (i) Data Pre-processing, (ii)

Find temporal frequent itemsets (iii) Identify temporal association rules and Generate rules set and output.

Data preprocessing performs several steps to improve the quality of the input dataset. Some operations include removal of unwanted or irrelevant data, integration of databases and data exchange and data reduction. The second step uses time constraints on the two parameters support and confidence to general frequent itemsets. Using these frequent itemsets, the next step generates association rules. Time constraint can be ignored at this stage, as it is already included during frequent itemset generation step. The resulting rules are called temporal rules. Out of this, the second and third steps have received much attention.

The temporal association rules introduced in Ale and Rossi (2000) are an extension of the non-temporal model. The basic idea is to limit the search for frequent sets of items, or itemsets, to the lifetime of the itemset's members. On the other hand, to avoid considering frequent an itemset with a very short period of life (for example, an item that is sold once), the concept of temporal support is introduced. Thus, each rule has an associated time frame, corresponding to the lifetime of the items participating in the rule. If the extent of a rule's lifetime exceeds a minimum stipulated by the user, we analyze whether the rule is frequent in that period. This concept allows us to find rules that with the traditional frequency viewpoint, it would not be possible to discover.

The lifespan of an itemset may include a set of subintervals. The subintervals are those such that the given itemset: (a) has maximal temporal support and (b) is frequent. This new model addresses the solution of two problems: (1) Itemsets not frequent in the entire lifespan but just in certain subintervals, and (2) the discovery of every itemset frequent in, at least, subintervals resulting from the intersection of the lifespans of their components, assuring in this way the anti-monotone property (Agrawal and Srikant, 1994).

Han *et al*. (2000a) proposed an algorithm that avoided the candidate generation of apriori algorithm for frequent pattern mining. This algorithm was later extended for sequential data by (Han *et al.*, 2000b). According to Das *et al*. (1998), extending the conventional association rule ($A \Rightarrow B$ – if A occurs then B occurs) to include the time characteristic ($A \Rightarrow^{T} B$ – if A occurs then B occurs within a time T) will produce quality frequent itemsets. However, such algorithms, normally require a new definition of support and confidence to include the new time characteristic included.

Chang et al. (2002) explore a new model of mining general temporal association rules from large databases where the exhibition periods of the items are allowed to be different from one to another. The algorithm referred to as algorithm SPF (Segmented Progressive Filter) first segments the database into sub-databases in such a way that items in each sub-database will have either the common starting time or the common ending time. Then, for each sub-database, SPF progressively filters candidate 2-itemsets with cumulative filtering thresholds either forward or backward in time. This feature allows SPF of adopting the scan reduction technique by generating all candidate k-itemsets ($k > 2$) from candidate 2-itemsets directly. The experimental results showed that algorithm SPF significantly outperforms other schemes which are extended from prior methods in terms of the execution time and scalability.

Tansel and Imberman (2007) proposed a method where association rules were extracted for consecutive time intervals with different time granularities. An enumeration operation that extracts portions of a temporal relation was used during mining process and was combined with the first step of discovering association rules. Using this approach, the process of knowledge discovery can observe the changes and fluctuation in the association rules over the time period when these rules are valid.

Thuan (2010) extended the a priori algorithm and developed an optimization technique for mining time pattern association rules. The algorithm, termed as, Mining of Time Pattern Association (MTPA) rules algorithm, identifies all rules that are repeated daily, monthly, yearly, thus exploiting the temporal characteristic. The report provides proof for the correctness of MTPA but lacks in experimental results.

Gharib et al. (2010) presented a system for generating temporal association rules to solve the problem of handling time series by including time expressions into association rules. They considered the problem of reworking when new records are added into the temporal database. To solve this they extended an incremental algorithm to maintain the temporal association rules in a transaction database, at the same time maintains the benefits from the results of earlier mining to derive the final mining output. The experimental results on both the synthetic and the real dataset illustrated a significant improvement over the conventional approach of mining the entire updated database.

An algorithm termed as T-Apriori was proposed by Liang *et al*. (2005) which modified the traditional Apriori algorithm to include temporal characteristics. The efficiency of extracting temporal association rules was demonstrated using an ecological database. However, performance evaluation of the algorithm is not presented.

Ning et al. (2006) proposed an temporal association rule mining algorithm, where the order of mining steps were altered. In the first step, the algorithm generated association rules were created in the traditional fashion, after generation the valid time constraints were considered. Because there are no time-constraints, it is unavoidable to scan the database repeatedly. The candidate frequent itemset generated by the algorithm promptly decreases. It is unnecessary to take mergence mining operations. The time complexity of the execution of the first step is negligible.

Using a non Apriori-based technique that avoids multiple database scans, (Winarko and Roddick, 2005) achieved to efficiently mine arrangements and rules in a temporal database. However, in their methods they do not consider any constraints for the temporal relations and do not examine any measures for their rules other than the traditional confidence.

4. CONCLUSION

Analyzing large sequential data streams to unearth any hidden regularities is important in many applications ranging from finance to manufacturing processes to bioinformatics. In this article, an overview of temporal data mining techniques for such problems is presented. From the study, it is understood that the field is not yet considered mature and improvements in many areas can be considered. In particular, three areas can be explored. The first is to analyze methods that can reduce number of passes over the database. This will reduce the number of computations required during the generation of frequent patterns and association rules. The second is to add extra constraint on the structure of the patterns. Identifying and removing redundant temporal rules can also be analyzed. These challenges will be considered in future and experiments to evaluate these methods will be proposed and implemented.

REFERENCES

[1] Agrawal, R. and Srikant, R. (1994) Fast algorithms for mining association rules in large databases, Proceedings of 20th International Conference on Very Large Data Bases, Pp 487–499.

[2] Agrawal, R. and Srikant, R. (1995) Mining sequential patterns, in: P.S. Yu, A.S.P. Chen (Eds.), Proceedings of the 11th International Conference on Data Engineering (ICDE'95), IEE Computer Society Press, Taipei, Taiwan, Pp. 3–14.

[3] Agrawal, R., Imielinski, T. and Swami, A.N. (1993) A Mining association rules between sets of items in large databases, Proc. ACM SIGMOD Conf. on Management of Data, pp 207–216.

[4] Ale, J.M. and Rossi, G.H. (2000) An approach to discovering temporal association rules, Proceedings of the 2000 ACM Symposium on Applied Computing, Pp. 294–300.

[5] Antunes, C. M. & Oliveira, A. L. (2001), Temporal data mining: An overview, 'Proceedings of the KDD'01 Workshop on Temporal Data Mining', San Francisco, USA, pp. 1–13.

[6] Chang, C.Y., Chen, M.S. and Lee, C.H. (2002) Mining General Temporal Association Rules for Items with Different Exhibition Periods, Proceedings of the 2002 IEEE International Conference on Data Mining (ICDM'02)IEEE Computer Society,Maebashi City, Japan, Pp.59-66.

[7] Das, G., Mannila, H. and Smyth, P. (1998) Rule Discovery from Time Series. KDD, Pp.16-22

[8] Dunham, M. (2003) Data Mining: Introductory and Advanced Topics, Prentice Hall.

[9] Gadia, S.K. and Nair, S.S. (1993) Temporal Databases: A Prelude to Parametic Data, Temporal Databases: Theory, Design, and Implementation, eds. A. Tansel, et al., The Benjamin/Cummings Publishing Company, Inc., Pp.28-66.

[10] Gharib, T.F., Nassar, H., Taha. M. and Abraham, A. (2010) An efficient algorithm for incremental mining of temporal association rules, Data and Knowledge Engineering, Vol. 69, Pp. 800-815.

[11] Han, J. and Kamber, M. (2001) Data mining: Concepts and techniques, Morgan Kauffmann, San Fransisco, CA.

[12] Han, J., Pei, J. and Yin, Y. (2000a) Mining Frequent Patterns without Candidate Generation. ACM SIGMOD Int. Conf. on Management of Data, Pp. 1-12.

[13] Han, J., Pei, J., Mortazavi-Asl, B., Chen, Q., Dayal, U. and Hsu, M. (2000b) FreeSpan: Frequent pattern-projected sequential pattern mining. ACM SIGKDD, Pp.355-359.

[14] Hand, D., Mannila, H. and Smyth, P. (2001) Principles of data mining, MIT Press, Cambridge, MA

[15] Hipp, J., Guntzer, U. and Nakhaeizadeh, G. (2000) Algorithms for Association Rule Mining – A General Survey and Comparison, SIGKDD Explorations, Vol.2, No.2, Pp.1-58.

[16] Janetzko, D., Cherfi, H., Kennke, R., Napoli, A. and Toussaint, Y. (2004) Knowledge-based selection of association rules for text mining, Proceedings of ECAI'2004, IOS Press, Pp. 485–489.

[17] Laxman, S. and Sastry, P.S. (2006) A survey of temporal data mining, Academy Proceedings in Engineering Sciences, Vol. 31, No.2, Pp.173–198.

[18] Liang, Z., Xinming, T., Ling, L. and Webliang, J. (2005) Temporal association rule mining based on T-Apriori algorithm and its typical application, Proceedings of International Symposium on Spatio-temporal Modeling, Spatial Reasoning, Analysis, Data Mining and Data Fusion.

[19] Lu, H., Feng, L. and Han, J. (2000) Beyond intratransaction association analysis: mining multidimensional intertransaction association rules,

ACM Transactions on Information Systems, No. 18, Vol. 4, Pp. 423–454.

[20] Mannila, H. and Toivonen, H. (1996) Discovering generalised episodes using minimal occurrences, in: E. Simoudis, J. Han, U. Fayyad (Eds.), Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD'96), AAAI Press, Portland, Oregon, Pp. 146–151.

[21] Ning, H., Yuan, H. and Chen, S. (2006) Temporal Association Rules in Mining Method, Proceedings of the First International Multi-Symposiums on Computer and Computational Sciences (IMSCCS'06), IEEE Computer Society, Vol. 2, Pp.739-742.

[22] Ozden, B., Ramaswamy, S. & Silberschatz, A. (1998), Cyclic association rules, Proceedings of the 14th International Conference on Data Engineering (ICDE'98)', IEEE Computer Society Press, Orlando, Florida, USA, pp. 412–421.

[23] Roddick, J.F. and Spiliopoulou, M. (2002) Survey of temporal knowledge discovery paradigms and methods. IEEE Transactions on Knowledge and Data Engineering, Vol. 14, No. 4, Pp. 750–767.

[24] Tan, P.N., Kumar, V. and Srivastava, J. (2002) Selecting the right interestingness measure for association patterns, Proceedings of KDD'2002, Pp.32-41.

[25] Tansel, A.U. and Imberman, S.P (2007) Discovery of Association Rules in Temporal Databases, International Conference on Information Technology: New Generations - ITNG , pp. 371-376.

[26] Thuan, N.D. (2010) Mining Time Pattern Association Rules in Temporal Database, Journal of Communication and Computer, Vol. 7, No. 3 (Serial no. 64), Pp. 50-56.

[27] Winarko, E. and Roddick, J.F. (2005) Discovering richer temporal association rules from interval-based data, in A. M. Tjoa and J. Trujillo, eds, Proceedings of the 7th International Conference on Data Warehousing and Knowledge Discovery (DaWaK'05), Vol. 3589 of LNCS, Springer, Copenhagen, Denmark, Pp. 315–325.

[28] Witten, I. H. and Frank, E. (2000) Data mining: Practical machine learning tools and techniques with JAVA implementations, Morgan Kaufmann, San Fransisco, CA.