# TEMPORAL SEQUENTIAL PATTERN IN DATA MINING TASKS

DR. NAVEETA MEHTA
Asst. Prof., MMICT&BM, M.M. University, Mullana
navita80@gmail.com
MS. SHILPA DANG
Lecturer, MMICT&BM, M.M. University, Mullana
dangshilpa2@gmail.com

**Abstract**

The rapid increase in the data available leads to the difficulty for analyzing those data and different types of frameworks are required for unearthing useful knowledge that can be extracted from such databases. The field of temporal data mining is relatively young and one expects to see many new developments in the near future. In all data mining applications, the primary constraint is the large volume of data. Hence there is always a need for efficient algorithms.
This paper gives an overview of the temporal data mining task and highlights the related work in this context.
**Keywords:** Data Mining (DM); Knowledge Discovery in Database (KDD); Temporal Data Mining (TDM).

## 1. Introduction

A huge amount of data is collected everyday and a real universal challenge is to find actionable knowledge from such large amount of data. DM is an emerging research direction to meet this challenge. DM techniques can be deployed to search large databases to discover useful information that might otherwise remain unknown. Many data mining problems involve temporal aspects. Examples range from transaction databases in health care and insurance, stock exchange and customer goods in market sectors, sensor data collected from sensor networks, to scientific databases in geophysics and astronomy. Mining this temporal data poses interesting challenges than mining static data. While the analysis of static data sets often comes down to the question of relating data items, with temporal data there are many additional possible relations.
Rest of the paper is organized as follows. Section 2 describes temporal data mining, various data mining tasks like prediction, classification, clustering etc. w.r.t. temporal sequential pattern are explored in section 3. Section 4 describes the related work and finally article is concluded in Section 5.

## 2. Temporal Data Mining

Data mining is actually an integral part of Knowledge Discovery in Database (KDD) process, which is the overall process of converting raw data into useful information [1]. A typical KDD process which consists of five steps [2]:

1. Data collection and cleaning: selecting attributes, dealing with errors, identification of the necessary background knowledge, etc.
2. Choice of pattern discovery method: deciding on the types of knowledge to be Discovered, parameter selection, etc.
3. Discovery of patterns (data mining): running algorithms for discovering different types of patterns
4. Pattern Presentation: selecting interesting patterns, visualisation of results, etc.
5. Putting knowledge into use.

Temporal Data Mining (TDM) deals with the problem of mining patterns from temporal data, which can be either symbolic sequences or numerical time series. It has the capability to look for interesting correlations or rules in large sets of temporal data, which might be overlooked when the temporal component is ignored or treated as a simple numeric, attribute [3].Currently TDM is a fast expanding field with many research results reported and many new temporal data mining analysis methods or prototypes developed recently. There are two factors that contribute to the popularity of temporal data mining. The first factor is an increase in the volume of temporal data stored, as many real-world applications deal with huge amount of temporal data. The second factor is the mounting recognition in the value of temporal data.
In many application domains, temporal data are now being viewed as invaluable assets from which hidden knowledge can be derived, so as to help understand the past and/or plan for the future [4].

TDM covers a wide spectrum of paradigms for knowledge modelling and discovery. Since temporal data mining is relatively a new field of research, there is no widely accepted taxonomy yet. Several approaches have been used to classify data mining problems and algorithms. Roddick & Spiliopoulou (2002) [3] 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.

3. Temporal Sequential Pattern in Data Mining Tasks

Data mining has been used in a wide range of applications. However, the possible objectives of data mining, which are often called tasks of data mining, can be classified into some broad categories: prediction, classification, clustering, search and retrieval, and pattern discovery [5]. This categorization follows the categorization of data mining tasks extended to temporal data mining [6].

***3.1. Prediction:*** Prediction is the task of explicitly modelling variable dependencies to predict a subset of the variables from others. The task of time series prediction is to forecast future values of the time series based on its past samples. In order to perform the prediction, one needs to build a predictive model from the data. Koskela et al. (1996) [7] have studied neural networks for nonlinear modelling of time series data. The prediction problem for symbolic sequences has been addressed in AI research by Dietterich and Michalski (1985) [8].

***3.2. Classification:*** Classification is the task of assigning class labels to the data according to a model learned from the training data where the classes are known. Classification is one of the most common tasks in supervised learning, but it has not received much attention in temporal data mining [9]. In sequence classification, each sequence presented to the system is assumed to belong to one of predefined classes and the goal is to automatically determine the corresponding category for a given input sequence. Examples of sequence classification applications include signature verification [10], gesture recognition [11], and hand-written word recognition [12].

***3.3. Clustering:*** Clustering is the process of finding intrinsic groups, called clusters, in the data. Clustering of time series is concerned with grouping a collection of time series (or sequences) based on their similarity. Time series clustering has been shown effective in providing useful information in various domains [13]. For example, in financial data, clustering can be used to group stocks that exhibit similar trends in price movements. Goutte et.al. (1999) identifying clustering of fMRI time series for identifying regions with similar patterns of activation [14]. Clustering of sequences is relatively less explored but is becoming increasingly important in data mining applications such as web usage mining and bioinformatics [5]. A survey on clustering time series has been presented by Liao (2005) [13].

***3.4. Searching and Retrieval:*** Searching and retrieval are concerned with efficiently locating subsequences or sub-series in large databases of sequences or time series. In data mining, query based searches are more concerned with the problem of efficiently locating approximate matching than exact matching, known as content-based retrieval. An example of a time series retrieval application is to find out all the days of the year in which a particular stock had similar movements to those of today. Another example is finding products with similar demand cycles.

An example of sequence retrieval is finding gene expression patterns that are similar to the expression pattern of a given gene. In order to address the time series retrieval problem, different notions of similarity between time series and indexing techniques have been proposed. There is considerably less work in the area of sequence retrieval, and the problem is more general and difficult. For more detail about time series and sequence retrieval can be found in Das and Gunopulos (2003) [15].

***3.5. Pattern Discovery:*** Unlike in search and retrieval applications, in the pattern discovery there is no specific query in hand with which to search the database. The objective is simply to discover all patterns of interest. While the other tasks described earlier have their origins in other disciplines like statistics, machine learning or pattern recognition, the pattern discovery task has its origin in data mining itself.

A pattern is a local structure in the data. There are many ways of defining what constitutes a pattern. There is no universal notion for interestingness of a pattern either. However, one concept that is normally used in data mining is that of frequent patterns, that is, patterns that occurs many times in the data. Much of data mining literature is concerned with formulating useful pattern structures and developing efficient algorithms for discovering frequent patterns.

Methods for finding frequent patterns are important because they can be used for discovering useful rules, which in turn can be used to infer some interesting regularities in the data. A rule usually consists of a pair of a left-hand side proposition (the antecedent) and a right-hand side proposition (the consequent). The rule states that when the antecedent is true, then the consequent will be true as well.

Therefore, to apply the pattern discovery methods on time series data, the time series should be first converted into a discrete representation, for example by first forming subsequences (using a sliding window) and then clustering these subsequence's using a suitable measure of pattern similarity [16].Another method can be used by quantizing the time series into levels and representing each level (e.g., high, medium, etc.) by a symbol [17]. A survey on time series abstraction methods can be found in Hoppner (2002) [18].

## 4. Related Work

A sequence database consists of ordered elements or events, recorded with or without a concrete notion of time. There are many applications involving sequence data, such as customer shopping sequences, Web click streams, and biological sequences. Mining sequential patterns is an active data mining domain dedicated to sequential data. For example, customer purchases, Web log access, DNA sequences, geophysical data, and so on. The objective is to find all patterns satisfying some given criterion that can be hidden within a set of event sequences. Sequential pattern mining, the mining of frequently occurring ordered events or subsequence as patterns, was first introduced by Agrawal and Srikant (1995) [19].

Generalized Sequential Patterns (GSP), a representative Apriori-based sequential pattern mining algorithm, proposed by Srikant and Agrawal (1996) [20], uses the downward-closure property of sequential patterns and adopts a multiple pass, candidate generate-and-test approach. GSP also generalized their earlier notion in Agrawal and Srikant (1995) [19] to include time constraints, a sliding time window, and user-defined taxonomies.

The studies of sequential pattern mining have been extended in several different ways. Mannila et al. (1997) [21] consider frequent episodes in sequences, where episodes are essentially acyclic graphs of events whose edges specify the temporal before-and-after relationship but without timing-interval restrictions. Sequence pattern mining for plan failures was proposed in Zaki et al. (1998) [22]. Garofalakis et al. (1999) [23] proposed the use of regular expressions as a flexible constraint specification tool that enables user-controlled focus to be incorporated into the sequential pattern mining process. The embedding of multidimensional, multilevel information into a transformed sequence database for sequential pattern mining was proposed by Pinto et al. (2001) [24]. Pei et al. (2002) [25] studied issues regarding constraint-based sequential pattern mining. CLUSEQ is a sequence clustering algorithm, developed by Yang and Wang (2003) [26]. An incremental sequential pattern mining algorithm, IncSpan, was proposed by Cheng et al. (2004) [27]. SeqIndex, efficient sequence indexing by frequent and discriminative analysis of sequential patterns, was studied by Cheng et al. (2005) [28].

Zaki (2001) [29] developed a vertical format-based sequential pattern mining method called SPADE, which is an extension of vertical format-based frequent item set mining methods. Prefix Span, a pattern-growth approach to sequential pattern mining, was developed by Pei et al. (2001) [30]. Prefix Span works in a divide-and-conquer way. The first scan of the database derives the set of length-1 sequential patterns. Each sequential pattern is treated as a prefix and the complete set of sequential patterns can be partitioned into different subsets according to different prefixes. To mine the subsets of sequential patterns, corresponding projected databases are constructed and mined recursively.

The CloSpan algorithm for mining closed sequential patterns was proposed by Yan et al. (2003) [31]. The method is based on a property of sequence databases, called equivalence of projected databases. CloSpan can prune the non-closed sequences from further consideration during the mining process. A later algorithm called BIDE, a bidirectional search for mining frequent closed sequences was developed by Wang and Han (2004) [32], which can further optimize this process by projecting sequence datasets in two directions.

A method for parallel mining of closed sequential patterns was proposed by Cong et al. (2005) [33]. A method, MSPX, for mining maximal sequential patterns by using multiple samples, was proposed by Luo and Chung (2005) [34]. Data mining for periodicity analysis has been an interesting theme in data mining. The notion of mining partial periodicity was first proposed by Han, Dong, and Yin, together with a max-sub pattern hit set method [35]. Ma and Hellerstein (2001) [36] proposed a method for mining partially periodic event patterns with unknown periods. Yang et al. (2003) [37] studied mining asynchronous periodic patterns in time-series data. Mining partial order from unordered data was studied by Gionis et al. (2003) [38] and Ukkonen et al. (2005) [39]. Pei et al. (2005) [40] proposed an algorithm for mining frequent closed partial orders from string sequences.

## 5. CONCLUSION

Data Mining is actually useful for designers, researchers and analysts for gaining knowledge from active and passive databases. Temporal data mining is one of the active statistical techniques for hidden temporal patterns. Temporal data mining is a very fast expanding field with many new research results reported and many new temporal data mining analysis methods or prototypes developed recently. In this paper temporal sequential

pattern based data mining tasks and a number of areas which are related to Temporal Data Mining in their objectives are explored.

**References:**


[1] U.M. Fayyad, G. Piatetsky-Shapiro and P. Smyth (1996): "The KDD process for Extracting useful knowledge from volumes of data", Communication of the ACM, Vol. 39, No. 11, pp. 27–34.

[2] M. Klemettinen, H. Mannila and H. Toivonen (1996): "Interactive exploration of discovered knowledge", a methodology for interaction, and usability studies, Technical Report C-1996-3, Department of Computer Science.

[3] J.F. Roddick and M. Spiliopoulou (2002): "A survey of temporal knowledge discovery paradigms and methods", IEEE Transactions on Knowledge and Data Engineering Vol.14, No.4, pp. 750–767.

[4] X. Chen and I. Petrounias (1998): "A framework for temporal data mining", in Proceedings of the 9th International Conference on Database and Expert Systems Applications (DEXA'98)', Vienna, Austria, pp. 796–805.

[5] S. Laxman and P.S. Sastry (2006): "A survey of temporal data mining", Vol. 31, pp. 173–198.

[6] J. Han and M. Kamber (2001): "Data Mining: Concepts and Techniques", Academic Press, San Diego.

[7] T. Koskela, M. Lehtokangas, J. Saarinem and K. Kaski (1996): "Time series prediction with multilayer perceptron, FIR, and Elman neural networks", in Proceedings of World Congress on Neural Network, pp. 491–496.

[8] T.G. Dietterich and R.S. Michalski (1985): "Discovering patterns in sequences of events", Artificial Intelligent, Vol.25, No. 2, pp. 187–232.

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

[10] V.S. Nalwa (1997): "Automatic on-line signature verification", Proceedings of the IEEE 85, pp. 215–239.

[11] J. Yamato, J. Ohya and K. Ishii (1992): "Recognizing human action in time sequential images using Hidden Markov Model", in Proceedings of 1992 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'92), pp. 379–385.

[12] A. Kundu, Y. He and P. Bahl (1988): "Word recognition and word hypothesis generation for handwritten script: A Hidden Markov Model based approach", in Proceedings of 1988 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'88), pp. 457–462.

[13] T. W. Liao (2005): "Clustering of time series data - a survey", Pattern Recognition Vol. 38, No.11, pp. 1857–1874.

[14] C. Goutte, P. Toft and E. Rostrup (1999): "On clustering fMRI time series", Vol. 9, No.3, pp. 298–310.

[15] G. Das and D. Gunopulos (2003): "Time series similarity and indexing", Invited Chapter in Handbook on Data Mining.

[16] G. Das, K. Lin, H. Mannila, G. Renganathan and P. Smyth (1998): " Rule discovery from time series" , in Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining, USA, pp. 16–22 .

[17] W. G. Aref, M. G. Elfeky and A. K Elmagarmid (2004): "Incremental, online, and merge mining of partial periodic patterns in time-series databases", IEEE Transactions on Knowledge and Data Engineering Vol. 16, No.3, pp. 332–342.

[18] F. Hoppner (2002) : " Time series abstraction methods - a survey", in Proceedings GI Jahrestagung Informatik Workshop on Knowledge Discovery in Databases, Lecture Notes in Informatics, Dortmund, Germany, pp. 777– 786.

[19] R. Agrawal, R. Srikant (1995): "Mining sequential patterns", In Proceedings of the international conference on data engineering (ICDE'95), Taipei, Taiwan, pp. 3–14.

[20] R. Srikant, R. Agrawal (1996): "Mining sequential patterns: generalizations and performance improvements", In Proceeding of the 5th international conference on extending database technology (EDBT'96), Avignon, France, pp. 3–17.

[21] H. Mannila, H. Toivonen, AI Verkamo (1997): "Discovery of frequent episodes in event sequences", Data Min Knowl Discov 1, pp. 259–289.

[22] MJ. Zaki (1998): "Efficient enumeration of frequent sequences", In Proceeding of the 7th international conference on information and knowledge management (CIKM'98), pp. 68–75.

[23] M. Garofalakis, R. Rastogi, K. Shim (1999): "SPIRIT: Sequential pattern mining with regular expression constraints", In Proceeding of the 1999 international conference on Very large data bases (VLDB'99), pp. 223–234.

[24] H. Pinto, J. Han, J. Pei, K. Wang, Q. Chen, U. Dayal (2001): "Multi-dimensional sequential pattern mining", In Proceeding of the 2001 international conference on information and knowledge management (CIKM'01), pp. 81–88.

[25] J. Pei, J. Han, W. Wang (2002) : " Constraint-based sequential pattern mining in large databases" , " In Proceeding of the 2002 international conference on information and knowledge management (CIKM'02), pp. 18–25.

[26] J. Yang, W. Wang (2003): "CLUSEQ: efficient and effective sequence clustering. In: Proceeding of the 2003 international conference on data engineering (ICDE'03), Bangalore, India, pp. 101–112.

[27] H. Cheng, X. Yan, J. Han (2004): "IncSpan: incremental mining of sequential patterns in large", In Proceeding of the 2004 ACM SIGKDD international conference on knowledge discovery in databases (KDD'04), pp. 527–532.

[28] H. Cheng, X. Yan, J. Han (2005): "Seqindex: indexing sequences by sequential pattern analysis", In Proceeding of the international conference on data mining (SDM'05), pp. 601–605.

[29] M. Zaki (2001): "SPADE: an efficient algorithm for mining frequent sequences", pp.31–60.

[30] J. Pei, J. Han, B. Mortazavil, H. Pinto, Q. Chen, U. Dayal, C.Hsu (2001): " Prefix Span: mining sequential patterns efficiently by prefix-projected pattern growth", In Proceeding of the 2001 international conference on data engineering (ICDE'01), Heidelberg, Germany, pp. 215–224.

[31] X. Yan, J. Han (2003): "Close Graph: mining closed frequent graph patterns", In Proceeding of the international conference on knowledge discovery and data mining (KDD'03), Washington, pp. 286–295.

[32] J. Wang, J. Han (2004): "BIDE: Efficient mining of frequent closed sequences", In Proceeding of the 2004 international conference on data engineering (ICDE'04), pp. 79–90.

[33] S.Cong, J. Han, D. Padua (2005): "Parallel mining of closed sequential patterns", In Proceeding of the international conference on knowledge discovery in databases (KDD'05), pp. 562–567.

[34] C. Luo, S. Chung (2005): "Efficient mining of maximal sequential patterns using multiple samples" , In Proceeding of the international conference on data mining (SDM'05), pp. 415–426.

[35] J. Han, G. Dong, Y. Yin (1999): "Efficient mining of partial periodic patterns in time series database", In Proceeding of the international conference on data engineering (ICDE'99), pp. 106–115.

[36] S. Ma, JL. Hellerstein (2001): "Mining partially periodic event patterns with unknown periods", In Proceeding of the international conference on data engineering (ICDE'01), pp. 205–214.

[37] J. Yang, W.Wang, PS.Yu (2003): "Mining asynchronous periodic patterns in time series data", IEEE Trans Knowl Data Eng 15, pp. 613–628.

[38] A. Gionis, T. Kujala, H. Mannila (2003): "Fragments of order ", In Proceeding of the international conference on knowledge discovery and data mining (KDD'03), pp. 129–136.
[39] A. Ukkonen, M. Fortelius, H. Mannila (2005): "Finding partial orders from unordered 0-1 data", In Proceeding of the international conference on knowledge discovery and data mining (KDD'05), pp. 285–293.
[40] J. Pei , J. Liu ,H. Wang ,K. Wang, PS.Yu ,J. Yang (2005): " Efficiently mining frequent closed partial orders", In Proceeding of the international conference on data mining (ICDM'05), pp. 753–756.