# Time Series Analysis and Mining Techniques

**Mehmet Sayal**
*Hewlett-Packard Labs, USA*

## INTRODUCTION

A *time series* is a sequence of data values that are recorded with equal or varying time intervals. Time series data usually includes timestamps that indicate the time at which each individual value in the times series is recorded. Time series data is usually transmitted in the form of a *data stream*, i.e., continuous flow of data values. Source of time series data can be any system that measures and records data values over the course of time. Some examples of time series data may be recorded from stock values, blood pressure of a patient, temperature of a room, amount of a product in the inventory, and amount of precipitation in a region. Proper analysis and mining of time series data may yield valuable knowledge about the underlying characteristics of the data source. Time series analysis and mining has applications in many domains, such as financial, biomedical, and meteorological applications, because time series data may be generated by various sources in different domains.

## BACKGROUND

Time series analysis and mining techniques differ in their goals and algorithms they use. Most of the existing techniques fall into one of the following categories:

- **Trend Analysis and Prediction:** The purpose is to predict the future values in a time series through analysis of historic values (Han & Kamber, 2001; Han, Pei, & Yin, 2000; Han et al., 2000; Kim, Lam, & Han, 2000; Pei, Tung, & Han 2001). For example, "How will the inventory amount change based on the historic data?" or "What will be the value of inventory amount next week?"
- **Similarity Search:** The most common purpose is to satisfy the user queries that search for *whole sequence* or *sub-sequence matching* among multiple time series data streams (Agrawal, Faloutsos, & Swami, 1993; Faloutsos, Ranganathan, & Manolopoulos, 1994; Kahveci & Singh, 2001; Kahveci, Singh, & Gurel, 2002; Popivanov & Miller, 2002; Wu, Agrawal, & El Abbadi, 2000; Zhu & Shasha, 2002). For example, "Can you find time series data streams that are similar to each other?" or "Which time series data streams repeat similar patterns every 2 hours?"
- **Relationship Analysis:** The main purpose is to identify relationships among multiple time series. Causal relationship is the most popular type, which detects the cause-effect relationships among multiple time series. For example, "Does an increase in price have any effect on profit?"

## MAIN THRUST

The techniques for predicting the future trend and values of time series data try to identify the following types of movements:

- Long-term or trend movements (Han & Kamber, 2001)
- *Seasonal and cyclic variations*, e.g., similar patterns that a time series appears to follow during corresponding months of successive years, or regular periods (Han & Kamber, 2001; Han, Pei, & Yin, 2000; Han, Pei, Mortazavi-Asl, et al., 2000; Pei, et al., 2001; Kim, et al., 2000)
- Random movements

Long-term trend analysis research is mostly dominated by application of well-studied statistical techniques, such as regression. Various statistical methods have been used for detecting seasonal and cyclic variations. Sequence mining techniques have also been used to detect repeating patterns. However, statistical methods are more suitable for detecting additive and multiplicative seasonality patterns in which the impact of seasonality adds up or multiplies the current values with each repetition. Random movements are usually considered as noise and ignored through the use of smoothing techniques.

Similarity search techniques have been studied in detail in the last ten years. Those techniques usually reduce the search space by extracting a few identifying features from time series data streams, and comparing the extracted features with each other to determine which time series data streams exhibit similar patterns. Some approaches look for whole pattern matching;

whereas, some others break the time series into segments and try to evaluate the similarity by comparing segments from different time series data streams. Most similarity search techniques use an indexing method in order to efficiently store and retrieve the feature sets that are extracted from time series data streams.

The general problem of similarity-based search is well known in the field of information retrieval, and many indexing methods exist to process queries efficiently. However, certain properties of time sequences make the standard indexing methods unsuitable. The fact that the values in the sequences usually are continuous, and that the elements may not be equally spaced in time dimension, makes it difficult to use standard text-indexing techniques like suffix-trees.

Faloutsos et al. introduced the most revolutionary ideas on similarity search (Agrawal, et al., 1993; Faloutsos, et al., 1994). Time series are converted into few features using *Discrete Fourier Transformation (DFT)* and indexed using *R-Trees* for fast retrieval. An important limitation of spatial indexing methods is that they work efficiently only when the number of dimensions is low. Therefore, the features extracted from time series data streams using DFT or any other method are not suitable for spatial indexing methods. The general solution to this problem is *dimensionality reduction*, i.e., to extract a signature of low dimensionality from the original feature set. The dimensionality reduction has to preserve the distances between the original data sets to some extent so that indexing and searching in the signature space can be done without loosing accuracy significantly. It was proven that *false dismissals* are avoided during dimensionality reduction, but *false alarms* are not avoided.

Several research papers applied similar approaches for transforming the time series data from time domain into frequency domain using DFT, while preserving the *Euclidean distance* among the original data sets to avoid false dismissals (Kahveci, et al., 2002; Zhu & Shasha, 2002). DFT provides a very efficient approximation for time series data streams, but it has its limitations too. For example, DFT preserves *Euclidean distance*, but loses phase information. Therefore, it is only possible to find out if a similarity exists between two or more time series data streams with DFT based techniques. It is not possible to tell anything about the time distance of similarity. There are some heuristic approaches trying to overcome this limitation, such as storing additional time information during transformation into frequency domain, but none of them seem to be very successful and they increase the complexity of algorithms. Discrete Wavelet Transformation (DWT) was also used in many research papers for feature extraction (Kahveci & Singh, 2001; Popivanov & Miller, 2002; Wu, et al., 2000). Those papers assumed that DWT was empirically superior to DFT, according to the results of a previous research. However, it was claimed later that such comparisons may be biased with regards to implementation details and selected parameters (Keogh & Kasetty, 2002).

Research on relationship analysis has recently started gaining momentum. The main purpose of a relationship analysis technique is to find out the relationships among multiple time series data streams. Causal relationships are the most common ones because discovery of causal relationships among time series data streams can be useful for many purposes, such as explaining why certain movements occur in the time series; finding out whether data values of one time series has any effect on the near-future values of any other time series; and predicting the future values of time series data stream not only based on its recent trend and fluctuations but also on the changes in data values of other time series data streams.

Early research papers on relationship analysis tried to make use of existing techniques from prediction and similarity search. Those approaches have certain limitations and new approaches are needed for relationship analysis. For example, prediction techniques consider the historic values of a time series and try to predict the future trend and fluctuations based on the historic trend and fluctuations. However, those techniques ignore the possibility that values in one time series may be affected by the values in many other time series. As another example, similarity search techniques can only tell whether two or more time series (or their segments) are similar to each other. Those techniques cannot provide details regarding the time domain when the impact of a change in the values of one time series is observed after a time delay on the values of another time series. This limitation occurs because the similarity model is different from the original data model in those techniques, i.e., data is transformed from time domain into frequency domain for enabling faster search, but certain features of the time series data, such as time relevant information, are lost in most of those techniques.

Recently introduced techniques can be applied in the time domain without having to transform data into another domain (Perng, Wang, Zhang, & Parker, 2000; Sayal, 2004). The main idea is to identify important data points in a time series that can be used as the characteristic features of the time series in time domain. The important points can be selected from the local extreme points (Perng, et al., 2000) or change points that correspond to the points in time where the trend of the data values in the time series has changed (Sayal, 2004).