Chapter 2.33
Improving Similarity Search in Time Series Using Wavelets

Ioannis Liabotis
University of Manchester, UK

Babis Theodoulidis
University of Manchester, UK

Mohamad Saraee
University of Salford, UK

ABSTRACT

Sequences constitute a large portion of data stored in databases. Data mining applications require the ability to process similarity queries over a large amount of time series data. The query processing performance is an important factor that needs to be taken into consideration. This article proposes a similarity retrieval algorithm for time series. The proposed approach utilizes wavelet transformation in order to reduce the dimensionality of the time series. The transformed series are indexed using X-Trees, which is a spatial indexing technique able to efficiently index high-dimensional data. The article proves that this technique outperforms the usage of the Fourier transformation, since the wavelet transformation provides better approximation of the time series. Through the experiments, it can be concluded that the optimum performance is obtained using 16 to 20 wavelet coefficients. Furthermore, a novel mechanism for reducing the complexity of the calculation for the false alarms removal is proposed. Storing the approximation coefficients of the penultimate level of the decomposition tree, the Euclidean distance between the two sequences is calculated, thus reducing further the number of false alarms before calculating the actual Euclidean distance using the complete time series. The article concludes with a detailed performance evaluation of the proposed similarity retrieval algorithm using data from the Greek stock market and the temperature measurements from Athens. The comparison is done with techniques that use the Haar transform and the R*-Tree, and the proposed algorithm is shown to outperform them.

INTRODUCTION

Time series are a sequence of numerical values usually recorded at regular intervals (e.g., secondly, daily, weekly, monthly, or yearly). Regularity is known to be the major element considered in

most of the times series. Although there are cases in which time series have no regularity (e.g., the history of stock splits), this work only considers time series with regularity. Time series appear in many database applications, from scientific to financial. Examples of such application domains include scientific experiments like temperature measurements over time; medical measurements, such as blood pressure or body temperature measurements taken in regular time intervals; business applications, including stock price indexes, such as opening, closing, minimum and maximum values or bank account histories, event sequences in automatic control, and musical or voice recordings.

Interesting features that can be retrieved from time series are similarities among them. There are various types of applications, depending on the type of sequence we are looking into. For example, a business analyst may desire to identify companies with similar patterns of growth to determine if they share other common characteristics. The identification of stocks with similar price movements could lead to predictions of their behavior. Also, similar patterns between two stocks that appear with a slight time difference can be identified. For example, the determination of time periods over the year that the daily temperature in Paris had the same fluctuation as January's daily temperature in Athens can help weather forecasting or prediction of unexpected weather conditions.

In this article, a similarity retrieval algorithm for time series is proposed. In the proposed algorithm, the wavelet transformation is used in order to reduce the dimensionality of the time series, and the transformed series are indexed using X-Trees (Berchtold et al., 1996). This indexing technique allows using a higher dimensionality for indexing without important loss in performance. From the evaluation process, it is shown that the wavelet transform reduces the false alarms significantly compared to the Fourier transform and provides better approximation of the time series. Also, using a number of dimensions between 16 and 20, the performance is optimized. Comparison has been done with techniques using the Haar transform and the R*-Tree (Chan & Fu, 1997; Wu et al., 2000; Chan et al., 2003), and the proposed algorithm has been shown to outperform them.

Furthermore, a novel mechanism for reducing the complexity of the false alarms removal calculation is proposed. Storing the approximation coefficients of the penultimate level of the decomposition tree, the Euclidean distance between the two sequences is calculated, thus reducing further the number of false alarms before calculating the actual Euclidean distance using the whole time series.

The organization of the rest of this article is as follows: the second section discusses the related work in relation to the work and assumptions reported in this article. The third section discusses the basics of the wavelets theory and the discrete wavelet transform. The similarity retrieval algorithm is discussed in the fourth section. The evaluation of the algorithm is presented in the fifth. Finally, the sixth section concludes the article.

RELATED WORK

In recent years, there has been a great deal of research in the similarity search area. There are several different similarity measurements that have been proposed in the literature as well as several algorithms for the efficient retrieval of these similarities. One approach to define the similarity of two time sequences is to use the Euclidean distance in an appropriate multidimensional space. In the work by Agrawal et al. (1993), two sequences are considered similar, if the Euclidean distance between them is less than a predefined threshold ε. The Euclidean distance between two sequences is defined as the square root of the sum of the squared differences: $D = (\Sigma (Xi-Yi)^2)^{1/2}$. The same distance metric is used