Recent Developments in Time Series Data Mining: Similarity Measures & Representations

Leonidas Karamitopoulos, Georgios Evangelidis

Dept. of Applied Informatics, University of Macedonia, Thessaloniki, Greece
Tel: +302310-891844, Email: {lkaramit, gevan}@uom.gr

In the last decade there has been an increasing interest in mining time series data since huge amounts are generated by several procedures in almost every domain such as in business, industry, medicine, science etc. Moreover, considering image or video data as time series data, the list of time series databases that need to be mined is expanded. During this period of time, hundreds of papers have been published covering all aspects of time series data mining, namely, dimensionality reduction or representation techniques, indexing, clustering, classification, novelty detection, motif discovery etc. Most of the contributions focus on proposing different dimensionality reduction approaches and providing novel similarity measures in order to deal with the unique characteristics of time series data, specifically, the high dimensionality, the high feature correlation and the large amounts of noise and to improve the performance of the existing data mining techniques. The objective of this paper is to serve as an overview of the most recent advances in the field of time series data mining. Although a general overview is included, the literature review is focused mainly on papers of the last three years.

1. Introduction

The Data Mining (DM) field has attracted a lot of attention during the last decade since it involves techniques and algorithms capable of efficiently extracting patterns that can potentially constitute knowledge from large databases. The primary goals of DM methods are the description of a particular dataset (often huge) and the prediction of future values of interest based on already known values from a database [13]. Time Series Data Mining (TSDM) is a relatively new field that is comprised by DM methods adjusted in a way that they take into consideration the temporal nature of data. Several procedures generate huge amounts of data in the form of time series, in almost every domain such as in business, industry, medicine and science. In addition to that, TSDM techniques can be applied on image or video data since these types of data can be considered also as time series. According to an electronic poll (126 voters) conducted in September 2005 by the kdnuggets site (http://www.kdnuggets.com), 40% of the voters stated that they analyzed/mined time series data during the previous 12 months. That was the second highest percentage in the corresponding sample (82% stated that they analyzed/mined "table data , fixed number of columns"). During the last decade hundreds of papers have been published covering all aspects of time series data mining. An indicative statistic of the increasing interest in mining time series arose while we were surveying the literature through the DBLP site (http://www.informatik.uni-trier.de/~ley/db/index.html). Providing the keyword "time series" for each one of the last 20 years, we monitored the number of papers, which included this term in their title. As

it is shown in Figure 1, this number increases almost exponentially for this period of time. Many of these papers are related to data mining tasks.

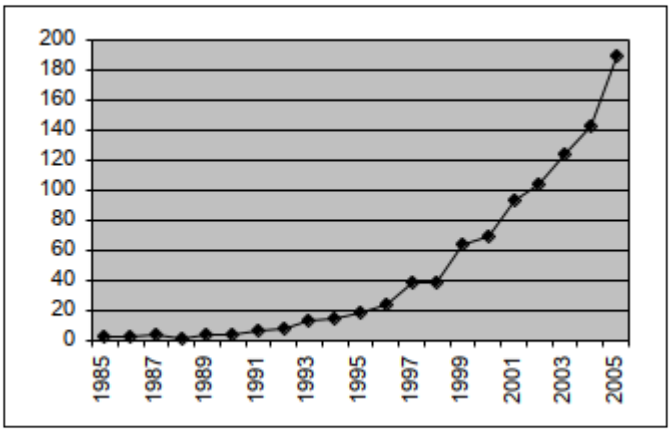

Figure 1: Number of papers including the term "Time Series" in their title.

According to the research in this field, the main tasks of TSDM methods are: forecasting, indexing, clustering, classification, novelty detection, motif discovery and rule discovery. Although some of these tasks are similar to the corresponding DM tasks, the temporal aspect arises some special issues to be considered and/or imposes some restrictions in the corresponding applications. First, in most of the above tasks it is necessary to define a similarity measure between two time series. A second issue that arises in TSDM and interrelates to the selection of a similarity measure is the representation of a time series. In addition, there have been introduced several applications that utilize existing techniques or modified versions of them. Through such application oriented research, application-specific representation schemes and similarity measures are defined that potentially may be of broader usefulness. As an example, a recent application is presented by Wu et al. [40] who propose a comprehensive approach to the problem of respiratory motion data in cancer radiation treatment. Although this is an application paper, a general framework is provided appropriate for analyzing any motion with structured time series data.

To our knowledge, there is no paper reviewing thoroughly the Time Series Data Mining field. However, there exist some excellent tutorials by Keogh [19] and Faloutsos [11] . In addition to these tutorials, Keogh and Kasetty [22] conducted a thorough survey on TSDM literature focusing on the empirical evaluations of the proposed approaches in the following tasks: indexing, clustering, classification and segmentation. Their experimental results suggested “the need for a set of time series benchmarks and more careful empirical evaluation in the data mining community”. There are two more papers that review the broader field of Temporal Knowledge Discovery where other types of data are also considered, along with time series, such as sequences of events. In the first paper, Roddick and Spiliopoulou [34] provide a framework for the categorization of studies on temporal data mining along three dimensions: (a) data-type (b) mining paradigm and (c) ordering. In the second paper, Antunes and Oliveira [3] provided another overview of Temporal Data Mining aiming also at the classification and organization of the available techniques. Their approach in discovering patterns from temporal data involves three steps: (a) the representation of the data, (b) the definition of similarity measures and (c) the application of data mining methods.

The objective of this paper is to serve as an overview of the most recent advances in the field of time series data mining focused mainly on papers of the last three years. The survey cannot be considered comprehensive since the emerging field of data mining has attracted the interest of researchers from many diverse fields such as computing science, bioinformatics, manufacturing etc. Moreover, the limited space of this paper led us to give more emphasis on the main contributions to time series representations and similarity measures, which consist the main research effort in this field.

In Section 2 we briefly provide the background of time series data mining tasks and techniques. Section 3 discusses various indexing issues. In Section 4, we present past and recent similarity measures. Section 5 describes the various types of time series representations proposed to date. Finally, a conclusion is presented in Section 6.

2. TSDM concepts and tasks

A time series is a collection of observations made sequentially through time. At each time point one or more measurements may be monitored corresponding to one or more attributes under consideration. The resulting time series is called univariate or multivariate respectively. In many cases the term sequence is used in order to refer to a time series, although some authors refer to this term only when the corresponding values are non-numerical. Throughout this paper the terms sequence and time series are being used interchangeably.

As mentioned in Section 1, the most common tasks of TSDM methods are: indexing, clustering, classification, novelty detection, motif discovery and rule discovery. In most of the cases, forecasting is based on the outcomes of the other tasks. A brief description of each task is given below.

Indexing: Find the most similar time series in a database to a given query time series.

Clustering: Find groups of time series in a database such that, time series of the same group are similar to each other whereas time series from different groups are dissimilar to each other.

Classification: Assign a given time series to a predefined group in a way that is more similar to other time series of the same group than it is to time series from other groups.

Novelty detection: Find all sections of a time series that contain a different behavior than the expected with respect to some base model.

Motif discovery: Detect previously unknown repeated patterns in a time series database.

Rule discovery: Infer rules from one or more time series describing the most possible behaviour that they might present at a specific time point (or interval).

The temporal aspect of data arises some special issues to be considered and/or imposes some restrictions in the corresponding applications. First, it is necessary to define a similarity measure between two time series and this issue is very important in TSDM since it involves a degree of subjectivity that might affect the final result. A lot of research has focused on defining different similarity measures in order to improve the performance of the corresponding methods. Second, it is necessary to

apply a representation scheme on the time series data. Since the amount of data may range from a few megabytes to terabytes, an appropriate representation of the time series is necessary in order to manipulate and analyze it efficiently. The desirable properties that this approach should hold are: (a) the completeness of feature extraction and (b) the reduction of the dimensionality "curse" [1]. More specifically, the method of extraction features should guarantee that there would be no pattern missed, the number of patterns falsely identified as interesting will be minimized and the dimensionality reduction will be substantial. In many cases also, the objective is to take advantage of the specific characteristics of a representation that make specific methods applicable (i.e. inducing rules, Markov models). Consequently, the majority of the researchers are focused on defining novel similarity measures and representation schemes in order to improve indexing performance. Past and recent work on these issues is presented in the next three sections (3,4 and 5).

Clustering and classification of time series rely heavily on the similarity measure and the representation scheme selected, thus, there are very few papers proposing a novel algorithm [32]. A recent survey on clustering time series is provided by Liao [27].

Novelty detection is a very important task in many areas. Several alternative terms for "novelty" have been used, such as, "anomaly", "interestingness", "surprising", "faults" to name a few. Moreover, many problems of finding periodic patterns can be considered as similar problems. The important point here is to provide a clear and concise definition of the corresponding notion. For instance, Keogh et al. [23] describe a pattern as surprising "if the frequency with which it appears, differs greatly from that expected given previous experience". The authors present a novel algorithm, called Tarzan, and provide useful pointers to relevant literature. Recently, Aref et al. [4] focus on discovering partial periodic patterns in one or more databases. They present algorithms for incremental mining (how to maintain discovered patterns over time as the database is being expanded) and online mining (how to perform changes in various thresholds of a mining task while it is in progress).

Motif discovery is a well-known task in the bioinformatics community [36] but only recently attracted the interest of the data mining community [29]. Motifs are defined to be previously unknown, frequently occurring patterns in a time series. These patterns may be of particular importance to other data mining tasks, such as, rule discovery and novelty detection. The recent work of Tanaka et al. [37] proposes a new method for identifying motifs from multi-dimensional time series. They apply Principal Component Analysis to reduce dimensionality and perform a symbolic representation. Then, the motif discovery procedure starts by calculating a description length of a pattern based on the "Minimum Description Length" principle.

Finally, rule discovery has been a major topic in the Data Mining literature, but again, within time series context the interest has been in the representation schemes selected in order to make specific methods applicable [10].

3. Indexing

Indexing approaches are mostly influenced by the pioneer work of Agrawal et al. [1], generalized by Faloutsos et al. [12]. The emerged framework from these papers, referred as GEMINI, can be summarized in the following steps [11]:

- extract k essential features from the time series
- map into a point in k-dimension feature space
- organize points with off-the-shelf spatial access method ('SAM')
- discard false alarms

The first and second step suggests the application of a representation scheme in order to reduce the dimensionality. However, this mapping should guarantee that it would return all the qualifying objects. This implies that the similarity measure in the k-dimension feature space should lower bound the corresponding similarity measure in the original space [28]. The third step is an opened selection, however most of the times R-tree structures are used. Other indexing structures may be vp-trees [7] [39], hB-trees and grid-files. The fourth step is a consequence of the fact that this approach can not guarantee that there will not be returned unqualified objects, thus these false alarms should be discarded in a post processing phase.

Recently, Vlachos et al. [38] presented an external memory indexing method for discovering similar multidimensional time series under time warping conditions. The main contribution of this work is the ability to support various distance measures without the need to reconstruct the index. Two approaches with respect to distance measures are taken under consideration, namely, the Longest Common Subsequence (LCS) and the Dynamic Time Warping (DTW). Their indexing technique works by splitting a set of multiple time series in multidimensional Minimum Bounding Rectangles (MBR) and storing them in an R-tree. For a given query, a Minimum Bounding Envelope (MBE) is constructed, that covers all the possible matching areas of the query under time warping conditions. This MBE is decomposed into MBRs and then probed in the R-tree index.

4. Time series representation

There have been several time series representations proposed in the literature, mainly on the purpose of reducing the intrinsically high dimensionality of time series. We will refer to some of the most commonly used representations. A hierarchy of various time series representations is presented in a tree diagram in [28]. Discrete Fourier Transform (DFT) [1] was one of the first representation schemes proposed within data mining context. DFT transforms a time series from the time domain into the frequency domain whereas a similar representation scheme, Discrete Wavelet Transform (DWT) [8], transforms it into the time/frequency or space/frequency domain. A comparison of these two representations is provided in [41]. Singular Value Decomposition (SVD) [25] performs a global transformation by rotating the axes of the entire dataset such that the first axis explains the maximum variance, the second axis explains the maximum of the remaining variance and is orthogonal to the first axis etc. Piecewise Aggregate Approximation (PAA) [43] [21] divides a time series into segments of equal length and records the mean of the corresponding values of each one. Adaptive Piecewise Constant Approximation (APCA) [20] is similar to PAA but allows segments of different lengths. Piecewise Linear Approximation (PLA) approximates a time series by a sequence of straight lines.

Recently, more representation schemes have been proposed in order to reduce dimensionality. The first class of these schemes consists of symbolic representations. Lin et al. [28] propose a Symbolic Aggregate Approximation (SAX)

method, which uses as a first step the PAA representation and then discretizes the transformed time series by using the properties of the normal probability distribution. The resulting “word” depends on a previously chosen alphabet size. Morchen and Ultsch [31] provide a new unsupervised discretization method of time series, called Persist, which results into a symbolic time series with symbols that retain their temporal aspect. This method requires that the time series does not change behavior fast and does not contain a long-term trend. The basic idea of this approach is that a time series is a sequence of states generated by an underlying process. Persist is based on the Kullback_Leibler divergence between the marginal and the self-transition probability distributions of the states (the discretization symbols) in order to discover these states. Taking into account the temporal order of the original data derives the resulting representation and that may be a useful feature in knowledge discovery, especially within rule-discovery and anomaly-detection tasks. Bagnal and Janacek [5] assess the affects of clipping original data on the clustering of time series. Each point of a series is mapped to 1 when it is above the population mean and to 0 when it is below. This representation is called clipping and has many advantages especially when the original series is long enough. It achieves adequate accuracy in clustering, it efficiently handles outliers and it provides the ability to employ algorithms developed for discrete or categorical data. The authors evaluate the effects of this representation on clustering. They fit ARMA models on the transformed data, and they conduct extensive experiments on different clustering methods, such as, k-means, k-medoids and hierarchical clustering. Megalooikonomou et al. [30] introduce a novel dimensionality reduction technique, called Piecewise Vector Quantized Approximation (PVQA). This technique is based on vector quantization that partitions each series into segments of equal length and uses vector quantization to represent each segment by the closest codeword from a codebook. The original time series is transformed to a lower dimensionality series of symbols. This approach requires a training phase in order to construct the codebook (the Generalized Lloyd Algorithm is applied), a data-encoding scheme and a distance measure.

A second class of representation schemes aims at transforming multivariate time series and / or streaming data. Papadimitriou et al. [33] introduce SPIRIT, a new approach of discovering patterns from streaming multiple time series. This approach identifies correlations and hidden variables among time series, in order to summarize the entire set of streams and provide useful means in efficient forecasting. The SPIRIT also satisfies the important requirements of an efficient streaming pattern discovery procedure, that is, it is streaming, it scales linearly with the number of time series, it is adaptive to changes and it is automatic. The basic method applied in order to identify correlations and hidden variables is the Principal Component Analysis. Cole et al. [9] provide a work that addresses the task of discovering correlated windows of time series (synchronously or with lags) over streaming data. They concentrate in the case where the time series are “uncooperative”, meaning that there does not exist a fundamental degree of regularity that would allow an efficient implementation of DFT or DWT transformations. The proposed method involves a combination of several techniques – sketches (random projections), convolution, structured random vectors, grid structures, and combinatorial design – in order to achieve high performance. Gionis and Mannila [17] introduce a different approach, which is mainly motivated from research on human genome sequences. However, this approach is more general and involves multivariate time series. The notion behind

their approach is that, the high variability that some time series very often exhibit, may be explained by the existence of several different sources that affect different segments of this series. More specifically, the task is to find a proper way to segment a time series into k segments, each of which comes from one of h different sources (k >>h). This task is analogous to clustering the points of a time series in h clusters with the additional constraint that a cluster may change at most k-1 times. Gionis and Mannila provide three algorithms for solving this problem and they test them on synthetic and genome data. In a recent paper, Fujimaki et al. [15] present a novel anomaly detection method for spacecrafts, based on Kernel Feature Space and directional distribution. Part of their work is to define an anomaly metric. Although this is an application-oriented method, the fact that it requires little a priori knowledge makes it potentially useful to other applications too.

Finally, Vlachos et al. [39] propose to represent a time series by applying discrete Fourier transformations and retain the k best Fourier coefficients instead of the first few ones. Although this paper is motivated by mining knowledge from the query logs of the MSN search engine, the proposed methods may be applied for time series data mining in general.

5. Similarity Measures

The definition of novel similarity measures has been one of the most researched areas in the TSDM field. Generally, they are strongly related to the representation scheme applied to the original data. However, there are some similarity measures that appear frequently in the literature. Most of the researchers' choices are based on the family of Lp norms, that include the Euclidean distance. Yi and Faloutsos [43] presented a novel and fast indexing scheme when the distance function is any of the arbitrary Lp norms ($p = 1, 2, \ldots, \infty$). Another similarity measure that attracted a lot of attention, Dynamic Time Warping (DTW), comes from the speech recognition field [6]. The main advantage of this measure is that it allows acceleration-deceleration of a series along the time dimension (nonlinear alignments are possible), however it is computationally expensive. Therefore, there has been much research in order to improve the incurred performance. Probabilistic approaches (probabilistic generative modeling) to measuring similarity were also proposed [24] [16]. Here, similarity between two sequences S and S´ is measured by calculating the likelihood that S´ is generated from a model, which was constructed from S. Markov models have been constructed and experimented. Another family of distance measures, Longest Common Subsequence Measures (LCS), often used in speech recognition and text pattern matching. As an example of this approach, we refer to the work of Agrawal et al. [2] who define two sequences as similar when they have enough, non-overlapping, time-ordered pairs of subsequences that are similar. Gunopoulos and Das [18] provide a thorough tutorial on the above distance measures along with measures not mentioned here and other TSDM related issues. Keogh and Kasetty [22] performed two classification tasks implementing 11 different distance measures. A surprising result was that, under those specific settings of the experiments, Euclidean distance outperformed the others with respect to the error rates.

Recently, the research on similarity measures is focused, mainly, on defining or adjusting distance measures for multivariate time series and streaming data, as well as, on improving existing measures such as DTW. Yang and Shahabi [42] introduce a similarity measure for Multivariate Time Series (MTS) datasets, called Eros, used

in k nearest neighbor (kNN) searches. Their proposed measure is based on Principal Component Analysis since it requires the extraction of the principal components and the associated eigenvalues from each MTS. Eros extends the Frobenius norm to measure the distance (weighted) between the two extracted matrices. Additionally, lower and upper bounds are obtained in order to satisfy triangle inequality. Li et al. [26] propose an algorithm for fast and efficient recognition of motions in multi-attribute continuous motion sequences. The main contribution of this paper is the definition of a similarity measure based on the analysis of Singular Value Decomposition (SVD) properties of similar multi-attribute motions. The proposed measure deals with noise and takes into account the different rates and durations of each motion. The authors also propose a five-phase algorithm for handling segmentation and recognition in real-time.

Sakurai et al. [35] propose the Fast search method for dynamic Time Warping (FTW) that satisfies the following criteria: (a) it is fast, (b) it produces no false dismissals, (c) it does not pose any restriction on the series length and (d) it supports for any, as well as for no restriction on warping scope. Their approach is based on a new lower bounding distance measure. They represent the sequence with approximate segments, not necessary of equal length, and operate on them. Three segments, the lower bound, the upper bound, and the time interval, correspond to each one of these approximate segments. In order to fulfill all of the above criteria, they provide algorithms for dynamic programming and searching adjusted to the properties of this representation. Fu et al. [14] propose a new technique to query time series that incorporates global scaling and time warping. The argument is that most real world problems require the ability to handle both types of distortion simultaneously. The approach is to scale the sequence by a bounded scaling factor and also to find nearest neighbor or evaluate range query by applying time warping. The authors provide definitions and proofs of the necessary lower bounds.

Furthermore, there is the expected contribution to defining similarity measures by papers that propose novel representation schemes, since these two tasks are interrelated to each other. For instance, some representation-specific measures are provided for the representations presented in [28] [30] [15].

6. Conclusion

We provided an overview of the recent advances in the field of time series data mining of the last three years. The rapidly increasing generation of time series data necessitates the development of new techniques and tools to analyze them efficiently and accurately. New data mining methods and algorithms should be developed to perform on that type of data taken into consideration their special characteristics. Two important issues emerge in the application of data mining techniques: the transformation of the original data to reduce dimensionality and the definition of a distance measure to identify similar objects. This paper is mainly concentrated in the recent developments on these two issues.

There are three key observations. First, there is a trend to adjust existed techniques to data mining tasks performed on multivariate time series. The emerging need of data reduction leads the researchers in adopting well-known methods, such as principal component analysis, in their approaches. Second, the increasing demand of analyzing streaming data also results in modifying existing methods and

providing new algorithms. Third, there has been a lot of work in exploiting TSDM techniques from diverse application areas.

The fact that there has been a lot of research on the Time Series Data Mining field from several communities, besides computing science, suggests a closer, interdisciplinary cooperation. Future work will focus on broader literature review, covering advances from these communities, with the hope of bringing closer these research communities.

REFERENCES

[1] Agrawal R., Faloutsos C., Swami A., "Efficient Similarity Search In Sequence Databases", *Proc. FODO 1993*, pp.69-84, Chicago, Illinois (USA), October 1993.

[2] Agrawal R., Lin K.-I., Sawhney H. S., Shim K., "Fast Similarity Search in the Presence of Noise, Scaling, and Translation in Time-Series Databases", *Proc. VLDB 1995*, pp. 490-501, Zurich (Switzerland), September 1995.

[3] Antunes C.M., Oliveira A.L., "Temporal Data Mining: An Overview", *Workshop on Temporal Data Mining ACM SIGKDD 2001*, San Francisco, California (USA), August 2001.

[4] Aref W. G., Elfeky M. G., Elmagarmid A. K., "Incremental, Online, and Merge Mining of Partial Periodic Patterns in Time Series Databases", *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, no. 1, pp. 332-342, January 2004.

[5] Bagnall A. J., Janacek G. J., "Clustering Time Series from ARMA Models with Clipped Data", *Proc. ACM SIGKDD 2004*, pp. 49-58, Seattle, WA (USA), August 2004.

[6] Berndt D. J., Clifford J., "Using Dynamic Time Warping to Find Patterns in Time Series", *KDD Workshop 1994*, pp.359-370, Seattle, WA (USA), July 2004.

[7] Bozkaya T., Yazdani N., Ozsoyoglu M., "Matching and Indexing Sequences of Different Lengths", *Proc. CIKM 1997*, pp. 128-135, Las Vegas, Nevada (USA), November 1997.

[8] Chan K., Fu A. W., "Efficient Time Series Matching by Wavelets", *Proc. ICDE 1999*, pp. 126-133, Sydney (Australia), March 1999.

[9] Cole R., Shasha D., Zhao X., "Fast Window Correlations Over Uncooperative Time Series", *Proc. ACM SIGKDD 2005*, pp. 743-749, Chicago, Illinois (USA), August 2005.

[10] Das G., Lin K.I., Mannila H., Ranganathan G., Smyth P., "Rule Discovery from Time series", *Proc. ACM SIGKDD 1998*, pp. 16-22, New York, New York (USA), August 1998.

[11] Faloutsos C., "Mining Time Series Data", *Tutorial ICML 2003*, Washington DC (USA), August 2003.

[12] Faloutsos C., Ranganathan M., Manolopoulos Y., "Fast Subsequence Matching in Time Series Databases", *Proc. ACM SIGMOD 1994*, pp. 419-429, Minneapolis, Minnesota (USA), May 1994.

[13] Fayyad U.M., Piatetsky-Shapiro G., Smyth P., "From Data Mining to Knowledge Discovery: An Overview", *In Advances in knowledge Discovery and Data Mining*, pp. 1-30, AAAI / MIT Press 1996.

[14] Fu A. W, Keogh E, Lau L. Y. H., Ratanamahatana C., "Scaling and Time Warping in Time Series Querying", *Proc. VLDB 2005*, pp. 649-660, Trondheim (Norway), August-September 2005.

[15] Fujimaki R., Yairi T., Machida K., "An Approach to Spacecraft Anomaly Detection Problem Using Kernel Feature Space", *Proc. PAKDD 2005*, pp. 785-790, Hanoi (Vietnam), May 2005.

[16] Ge X., Smyth P., "Deformable Markov Model Templates for Time Series Pattern Matching", *Proc. ACM SIGKDD 2000*, pp. 81-90, Boston, Massachusetts (USA), August 2000.

[17] Gionis A., Mannila H., "Finding Recurrent Sources in Sequences", *Proc. RECOMB 2003*, pp. 123-130, Berlin Germany, April 2003.

[18] Gunopoulos D., Das G., "Time Series Similarity Measures", *Tutorial notes ACM SIGKDD 2000*, pp. 243-307, Boston, Massachusetts (USA), August 2000.

[19] Keogh E., "Data Mining and Machine Learning in Time Series Databases", *Tutorial ECML/PKDD 2003*, Cavtat-Dubrovnik (Croatia), September 2003.

[20] Keogh E., Chakrabarti K., Mehrotra S., Pazzani M., "Locally Adaptive Dimensionality Reduction for Indexing Large Time Series Databases", *Proc. ACM SIGMOD 2001*, pp. 151-162, Santa Barbara, California (USA), May 2001.

[21] Keogh E., Chakrabarti K., Pazzani M., Mehrotra S., “Dimensionality Reduction for Fast Similarity Search in Large Time Series Databases”, *Knowledge and Information Systems*, vol. 3, no. 3, pp. 263-286, February 2001.

[22] Keogh E., Kasetty S., “On the Need for Time Series Data Mining Benchmarks: A Survey and Empirical Demonstration”, *Proc. ACM SIGKDD 2002*, pp 102-111, Edmonton, Alberta (Canada), July 2002.

[23] Keogh E., Lonardi S., Chiu B. Y., “Finding surprising patterns in a time series database in linear time and space”, *Proc. ACM SIGKDD 2002*, pp. 550-556, Edmonton, Alberta (Canada), July 2002.

[24] Keogh E., Smyth P., “A Probabilistic Approach to Fast Pattern Matching in Time Series Databases”, *Proc. KDD 1997*, pp. 24-30, Newport Beach, California (USA), August 1997.

[25] Korn F., Jagadish H., Faloutsos C., “Efficiently Supporting Ad Hoc Queries in Large Datasets of Time Sequences”, *Proc. ACM SIGMOD 1997*, pp. 289–300, Tucson Arizona (USA), June 1997.

[26] Li C., Zhai P., Zheng S. Q., Prabhakaran B., “Segmentation and Recognition of Multi-Attribute Motion Sequences”, *Proc. ACM Multimedia 2004*, pp. 836-843, New York, New York (USA), October 2004.

[27] Liao T. W. (2005), “Clustering of time series data – a survey”, Pattern Recognition, 38 1857 1874

[28] Lin J., Keogh E., Lonardi S., Chiu B., “A Symbolic Representation of Time Series, with Implications for Streaming Algorithms”, *Proc. DMKD 2003*, pp. 2-11, San Diego California (USA), June 2003.

[29] Lin J., Keogh E., Lonardi S., Patel P., “Finding Motifs in Time Series”, *Proc. 2nd workshop on Temporal Data Mining ACK SIGKDD 2002*, pp. 53-68, Edmonton, Alberta (Canada), July 2002.

[30] Megalooikonomou V., Li G., Wang Q., “A Dimensionality Reduction Technique for Efficient Similarity Analysis of Time Series Databases”, *Proc. CIKM 2004*, pp. 160-161, Trondheim (Norway), August-September 2005.

[31] Morchen F., Ultsch A., “Optimizing Time Series Discretization for Knowledge Discovery”, *Proc. ACM SIGKDD 2005*, pp. 660-665, Chicago, Illinois (USA), August 2005.

[32] Oates T., “Identifying Distinctive Subsequences in Multivariate Time Series by Clustering”, *Proc.ACM SIGKDD 1999*, pp. 322-326, San Diego, California (USA), August 1999.

[33] Papadimitriou S., Sun J., Faloutsos C., “Streaming Pattern Discovery in Multiple Time Series”, *Proc. VLDB 2005*, pp. 697-708, Trondheim (Norway), August-September 2005.

[34] Roddick J.F., Spiliopoulou M., “A Survey of Temporal Knowledge Discovery Paradigms and Methods”, *IEEE Transactions On Knowledge And Data Engineering*, vol. 14, no. 4, pp. 750-767, July/August 2002.

[35] Sakurai Y., Yoshikawa M., Faloutsos C., “FTW: Fast Search under the Time Warping Distance”, *Pro. PODS 2005,* pp. 326-337, Baltimore, Maryland (USA), June 2005.

[36] Staden R., “Methods For Discovering Novel Motifs In Nucleic Acid Sequences”, *Computer Applications in Biosciences*, vol. 5, no. 4, pp. 293-298, October 1989.

[37] Tanaka Y., Iwamoto K., Uehara K., “Discovery of Time Series Motif from Multi-Dimensional Data Based on MDL Principle”, *Machine Learning*, vol. 58, no. 2-3, pp. 269-300, February 2005.

[38] Vlachos M., Hadjieleftheriou M., Gunopoulos D., Keogh E., “Indexing Multi-Dimensional Time Series with Support for Multiple Distance Measures”, *Proc. ACM SIGKDD 2003*, pp. 216-225, Washington DC (USA), August 2003.

[39] Vlachos M., Meek C., Vagena Z., Gunopoulos D., “Identifying Similarities, Periodicities and Bursts for Online Search Queries”, *Proc. ACM SIGMOD 2004*, pp. 131-142, Paris (France), June 2004.

[40] Wu H., Salzberg B., Sharp G. C., Jiang S. B., Shirato H., Kaeli D. R., “Subsequence Matching on Structured Time Series Data”, *Proc.ACM SIGMOD 2005*, pp. 682-693, Baltimore, Maryland (USA), June 2005

[41] Wu Y. L., Agrawal D., Abbadi A. E., “A comparison of DFT and DWT Based Similarity Search in Time Series Databases”, *Proc. CIKM 2000*, pp. 488-495, McLean, Virginia (USA), November 2000.

[42] Yang K., Shahabi C., “A PCA-based Similarity Measure for Multivariate Time Series”, *Proc. MMDB 2004*, pp. 65-74, Arlington, Virginia (USA), November 2004.

[43] Yi B. K., Faloutsos C., “Fast Time Sequence Indexing for Arbitrary L_p Norms”,, *Proc. VLDB 2000*, pp. 385-394, Cairo (Egypt), September 2000.