CLUSTERING TIME SERIES ONLINE IN A TRANSFORMED SPACE

Hamid R. Arabnia, Junfeng Qu, Yinglei Song, Khaled Rasheed, Byron Jeff
United State of America
{hra, khaled}@cs.uga.edu, {jqu, bjeff@clayton.edu}, ysong@umes.edu

ABSTRACT

Similarity-based retrieval has attracted an increasing amount of attention in recent years. Although there are many different approaches, most are based on a common premise of dimensionality reduction and spatial access methods. Relative change of the time series data provides more meaning and insight view of problem domain.. This paper presents our efforts on considering the relative changes of time series during the time series matching process. A similarity distance measure that based on transformed difference space of a series of critical points is proposed. Based on experiments with financial time series data, it can be concluded that our distance measure works as good as the Euclidean distance measure based normalized data without any shifting and scaling and PAA approach. The distance measure proposed is a general distance metric and is suitable to deal with online similarity matching because it does not maintain stream statistics over data streams.

Keywords: data mining, time series, clustering, similarity matching, Euclidean distance.

1 INTRODUCTION

[1] Humans are good at telling the similarity between time series by just looking at their plots. Such knowledge must be encoded in the computer if we want to automate the detection of similarity among time series. In general, given any pair of time series, their similarity is usually measured by their correlation or distance. If we treat a time series as high dimensional points, which in time series it is, the Euclidean distance appears to be a natural choice for distance between time series. The Euclidean distance is defined as:

Given two time series sequence and with n=m, their Euclidean distance is defined as:

$$D(X,Y) \equiv \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (1)$$

We define that two sequences X and Y are in -match if D(X,Y) is less than or equal to . We define n-dimensional distance computation as the operation that computes the distance between two sequences of length n. Basically, there are essentially two ways the data might be organized[1].

- Whole sequence matching: In the whole sequence matching, all the time series that assumed to be compared are at the same length. The query time series q is of length n too. The Euclidean distance between the query time series and any time series to be compared with can be computed in linear time. Given a query threshold, the answer to a whole sequence similarity query for q is all the time series in the data set whose Euclidean distance with q are less than the threshold.
- Subsequence Matching: Here the time series in the data set can have different lengths. The lengths of these candidate time series are usually larger than the length of the query time series. The answer to a subsequence query is any subsequence of any candidate time series whose distance with q is less than .

Shasha and Zhu[2] pointed out that the Euclidean distance measure is not adequate as a

[1] An earlier version of the manuscript was published in the proceedings of the 2007 International Conference on Information and Knowledge Engineering(IKE'07: June 2007).

Hamid R. Arabnia is with the Department of Computer Science, the University of Georgia, Athens, GA 30602, USA (e-mail: hra@cs. uga.edu).
Junfeng Qu, corresponding author, is with the Department of Information Technology, Clayton State University, Morrow, GA 30260 USA (corresponding author to provide phone: 678-466-4406; e-mail: jqu@clayton.edu).
Yinglei Song is with the Department of Mathematics and Computer Science, University of Maryland at East Shore, Princess Anne, MD 21853 USA (e-mail: ysong@umes.edu)
Khaled Rasheed is with the Department of Computer Science, the University of Georgia, Athens, GA 30602 (e-mail: khaled@cs.uga.edu).
Byron Jeff is with the Department of Information Technology, Clayton State University, Morrow, GA 30260 (e-mail: byronjeff@clayton.edu).

flexible similarity measure between time series because:

- Two time series can be very similar even though they have different base lines or amplitude scales.
- The Euclidean distance between two time series of different lengths is undefined even though the time series are similar to each other.
- Two time series could be very similar even though they are not perfectly synchronized. The Euclidean distance that sums up the difference between each pair of corresponding data points between two time series is too rigid and will amplify the difference between time series.

In a given time series, the related change between two adjacent data points are often thought of where information is resident in. Especially in financial market data analysis, the amplitude difference is more important than the time difference. Therefore, transform the time series into space of difference because any similarity matching is more meaningful and provides more insight view of problem domain, especially in financial data analysis.

In this paper, we proposed our similarity measure on transformation of the original time series into a new series of critical change-points (which contains the difference information of original series and the similarity clustering is based on). The rest of paper is organized as follows. Section 2 discussed related works on time series similarity matching. Section 3 describes our distance measure on the transformed space. Section 4 includes our experimental test with our similarity distance measure on financial time series data. In section 5, we conclude our research and point out future research direction..

2 RELATED WORKS

There are two basic strategies to cope with high-dimensional problems. The first is simply to use a subset of relevant variables to construct the model. That is, to find a subset of p′ variables where p′<<p. The second is to transform the original p variables into a new set of p′ variables, where p′<<p.

There are many efforts have been done by researchers to reduce dimensionality of time series while increase performance of similarity matching. The original work by Agrawal et al. [3] utilizes the Discrete Fourier Transform (DFT) to perform the dimensionality reduction on the data, then use spatial access methods to index the data in the transformed space to speed up whole sequence similarity searching process. Faloutsos et al [4] extended the work of Agrawal further to perform subsequence similarity matching using a Discrete Fourier Transformation. The idea is to map each data sequence into a small set of multidimensional rectangles in feature space; then, these rectangles can be readily indexed using traditional spatial access methods, like R-tree. A sliding window over the data sequence was used to extract features. Moon et al[5] used a generalized windows to reduce false negatives from Faloutsos method due to lack of point-filtering effect. The general match method divides data sequences into generalized sliding windows and the query sequence into generalized disjoint windows to reduce false dismissal. The authors also proposed a method of estimating the optimal value of the sliding factor that minimizes the number of page access.

Because of efficiency of wavelet transform, it is also used as feature extraction functions. Chan and Wu[6] proposed to use Haar wavelet transform for time series indexing and showed that Euclidean distance is preserved in the Haar transformed domain and no false dismissal will occur. The research also showed that the Haar transform can outperform DFT, the method also accommodate vertical shift of time series. Gilbert et al [7] presented techniques for computing small space representation of massive data streams using wavelet-based approximations that consist of specific linear projections of underlying data. By capturing various linear projections of the data and using them to provide pointwise and rangesum estimation of data stream, the method only use small amount of space and per-item time as well as provide accurate representation of data. Huhtala et al[8] proposed using a wavelet transformation of a time series to produce a natural set of features for the sequence in order to mine similarities in aligned time series. The features generated by wavelet transformations describe properties of the time series both at various locations and at varying time granularities such that they are insensitive to changes in the vertical position, scaling, and overall trend of the time series. The authors also examined how the similarity between time series changes as a function of time or as a function of time granularity considered.

Wu et al[9] compared the feature vector extraction using Single Value Decomposition (SVD) and DFT. The results showed that the SVD overall outperforms DFT when query for a large number of neighbors or take a large radius. SVD also provided the best linear least squares error to data. Korn et al [10] proposed SVD with Deltas (SVDD) based on SVD algorithm that supports ad hoc queries in large time sequence datasets. The SVDD algorithm achieves excellent dimensionality reduction and only requires three passes over the dataset while preserving distances.

Piecewise Aggregate Approximation (PAA)[11] method was also proposed in the time series representation with dimensionality reduction. The PAA method reduced the data from *n* dimensions into *N* dimensions by dividing the time series data

into *N* equi-sized "frames". The mean value of the data falling within a frame is calculated and a vector of these values becomes the data reduced representation. In general, the transformation produces a piecewise constant approximation of the original sequence. Keogh et al [12] further modified the method and proposed adaptive piecewise constant approximation (APAC). APCA approximates each time series by a set of constant value segments of varying lengths such that their individual reconstruction errors are minimal. The authors showed how APCA can be indexed using a multidimensional index structure and proposed two distance measures in the indexed space that exploit the high fidelity of APCA for fast searching: a lower bounding Euclidean distance approximation, and a non-lower bounding, but very tight Euclidean distance approximation.

New distance measures are also explored by researchers such as dynamic time warping[[2, 13], longest common sequence[14]. Dynamic time warping algorithm allows stretching or squeezing time axis in the matching the similarity of time series. The method proposed by Das[14] takes into account outliers, different scaling functions, and variable sampling rate while measuring the similarity between two time series. Wu et al[15] proposed a comprehensive solution for analysis, clustering, and online prediction of respiratory motion using subsequence similarity matching. In the system, a motion signal is captured in real time as a data stream, and is analyzed immediately by a piecewise linear representation that generated from a finite state model to capture the relative importance of amplitude, frequency, and proximity in time.

There are extensive researches in the search of similarity in data streams. Some recent papers include[16-20]. When dealing with data stream, traditional methods for time series are inefficient. The dimensionality reduction methods are very costly but are applied only once. The index methods are static because the index is constructed only once. Therefore, in order to utilize the advantages provide by the traditional methods, the dimensionality reduction methods must be applied each time a value arrives and the index must up updated each time a value arrives. The research cited above adopts the incremental computation to reduce computation time needed, or incrementally maintain stream statistics over data streams.

3 METHODOLOGY

The current researches do not concern on relative position of corresponding end points in the time series in the selected distance measures. In this paper, the relative positions of end points of each time series were considered for similarity matching algorithm. In this paper a new distance measure, which address relative position of the corresponding data points is introduced and used for time similarity matching.

In order to compress the high-dimension time series data, a critical change-points are first extracted from the incoming data stream by the algorithm that uses dual-model approach[21]. The key is to define a series of critical change-points that can represent the time series without lost its original shape and structure. For the limitation of the paper, please refer paper[21] for details of the algorithm. Given a time series $T = \{(x_1, t_1), (x_2, t_2), ..., (x_n, t_n)\}$, the algorithm works by first choosing a small sample *S* by identifying a series of critical change-points from the data set, which are identified as 'o' in the fig. 1. The algorithm then maps all data points into the data set to points identified as critical points.

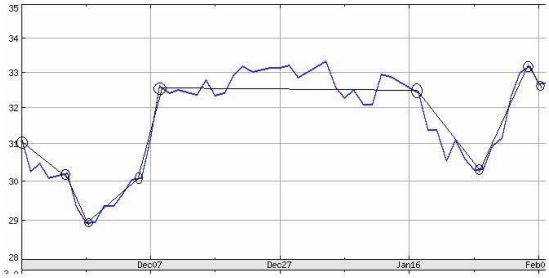

Figure 1. Critical change-points illustration

Specifically assume, without loss of generality, that $S = \{(s_1, t_1), (s_2, t_2), ..., (s_m, t_m)\}$, we define a mapping as follows:

We map s_1 to the origin and map the s_m to the end of data set, i.e. x_1 to s_1, and x_n to s_m. The size *m* of S that T mapped to completely depends on the shape of the T and the algorithm parameter chosen and in general $m << n$.

We further defined the similarity distance function that is based on the relative positions of the series of critical change-points discovered from time series as follows:

Define time series: $X = \{(x_1, t_1), (x_2, t_2), ..., (x_n, t_n)\}$ and $Y = \{(y_1, t_1), (y_2, t_2), ..., (y_n, t_n)\}$ which each has its sequence of critical change-points discovered so far $E_x = \{(e_1, t_1), (e_2, t_2), ..., (e_m, t_m)\}$ and $E_y = \{(e'_1, t'_1), (e'_2, t'_2), ..., (e'_m, t'_m)\}$, X and Y are similar if and only if they satisfy two conditions:

1. $d(E_x, E_y) < \gamma$

$d(E_x, E_y) =$

2. $$\sqrt{\sum_{i=1}^{m-1}\left[(e_{i+1}-e_i)-(e'_{i+1}-e'_i)\right]^2+\sum_{i=1}^{m-1}\left[(t_{i+1}-t_i)-(t'_{i+1}-t'_i)\right]^2}$$

where $\gamma \geq 0$ and is a user-defined parameter

In order to use metric distance indexing method for faster search, we need to show that our defined distance measure is a general distance metric. A general metric is a function δ that takes pairs of objects into real numbers, satisfying the following properties: for any objects *x, y, z*, $\delta(x, x) = 0$ and $\delta(x, y) > 0$, $x \neq y$ (non-negative definiteness); $\delta(x, y) = \delta(y, x)$ (symmetry); $\delta(x, y) \leq \delta(x, z) + \delta(z, y)$ (triangle inequality). Euclidean distance satisfies these properties.

For new sequence E_x and E_y (with the same length), the distance $d(E_x, E_y)$ is a general metric function.

To prove $d(E_x, E_y)$ is a general distance metric, we need to prove it is non-negative, symmetric, reflexive, and it satisfies the triangle inequality. Obviously, $d(E_x, E_y) \geq 0$ and $d(E_x, E_y) = d(E_y, E_x)$ from our definition, also $d(E_x, E_x) = 0$, so $d(E_x, E_y)$ is non-negative, symmetric and reflexive. Now we need to prove that $d(E_x, E_y)$ satisfies the triangle inequality, i.e. $d(E_x, E_y) \leq d(E_x, E_z) + d(E_z, E_y)$.

Given the sequences of events $E_x = \{(e_1, t_1), (e_2, t_2), ..., (e_m, t_m)\}$ and $E_y = \{(e'_1, t'_1), (e'_2, t'_2), ..., (e'_m, t'_m)\}$, we transform them into relative difference space of the sequence of events respectively as E_x to E and E_y into E' as:

$E = \{(\Delta e_1, \Delta t_1), (\Delta e_2, \Delta t_2), ..., (\Delta e_{m-1}, \Delta t_{m-1})\}$ and

$E' = \{(\Delta e'_1, \Delta t'_1), (\Delta e'_2, \Delta t'_2), ..., (\Delta e'_{m-1}, \Delta t'_{m-1})\}$

where $\Delta e_i = (e_{i+1} - e_i)$ and $\Delta e'_i = (e'_{i+1} - e'_i)$. Now it is obvious that triangle inequality is satisfied based on the Pythagorean theorem and Euclidean space.

After transforming the time series into series of critical change-points, how many of these critical change-points are necessary to represent the time series while retaining its structure and shape is critical. The compression ratio is defined to compare our approach with the PAA approach, which is different from ours and well studied[11].

The Compression ratio (CR) is defined as:

$$\frac{\text{Number of data points in the original time series}}{\text{Number of data points that represent the time series}}$$

.

4 EXPERIMENTS

We have proved that our new developed distance measure for time series similarity matches is a metric function. Shasha et al[2] showed that Euclidean distance alone does not give an intuitive measure of similarity under the conditions of the time series compared are of different baselines and scales. Therefore, shifting transform or scaling transforms are often performed before measure Euclidean distance. Here the shifting transform is defined as the transformation of old time series by adding some real number to each item into a new time series. Scaling transform on a time series is to get a new time series by multiplying some real number to each item in the old time series. A simple way to make a similarity measure invariant to shifting and scaling is to normalize the time series.

Define the normal form *Norm(X)* of a time series *X* is transformed from *X* by shifting the time series by its mean and then scaling by its standard deviation.

$$Norm(X) = \left(X - avg(X)\right) / std(X)$$

It is trivial that the normalized time series have the properties *avg(Norm(x))=0* and *std(Norm(X))=n* . The Euclidean distance between the normal forms of two time series is a similarity measure between time series that is invariant to shifting and scaling because they have the same baseline and scale[2].

we utilized seven stocks time series data (*EOG*, *SM*, *HAL*, *CDIS*, *NOVL*, *SCOX*, and *WMT*) from April 2005 to Oct.2005 to study our clustering measure,. For every possible pairing of the seven dataset from these stocks, we use group-average hierarchical clustering. The corresponding dendrogram of clustering based on different distance measure and transform techniques are shown.

We compared three distance measures:

1. Euclidean: The Euclidean distance measure as presented in the introduction is tested to facilitate comparison to the large body of literature that utilize this distance measure
2. PEuclidean: The PAA representation of time series with same compression ratio as our approach is also compared
3. DSDistance: The critical change-points representation of time series that are

clustered based on our defined transformed space

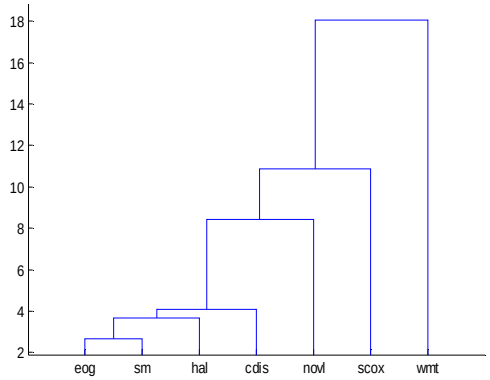

Figure 2. Euclidean distance cluster on normalized time series data

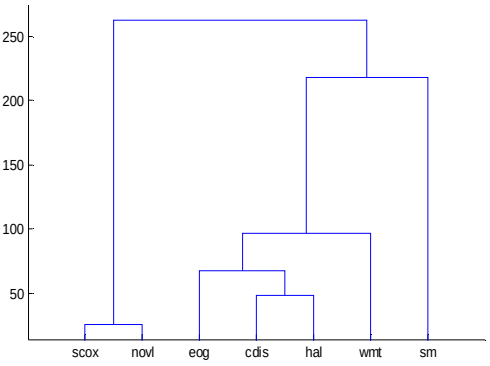

Figure 3. Euclidean distance cluster on raw time series data

Our critical change-point representation has the compression ratio around 25. Therefore, the PAA representation is compared with the same compression ratio. The clustering results are shown in fig. 4 and 5.

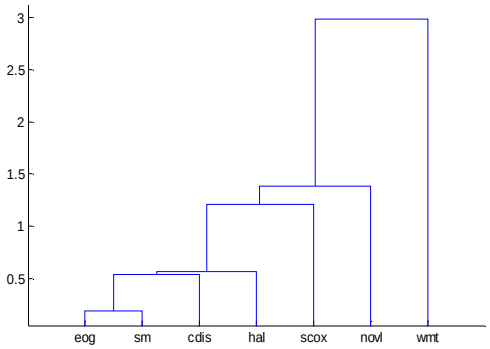

Figure 4. Euclidean distance cluster on normalized time series data on PAA representation (compression ratio=25)

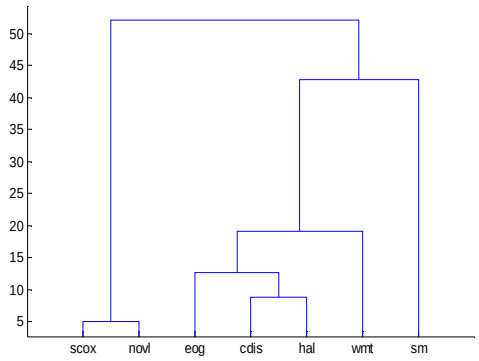

Figure 5. Euclidean distance cluster on raw time series data with PAA representation(compression ratio=25)

Figure 6 and 7 show our critical change-point presentation with compression ratio of 25. The distance measure on transformed space is used.

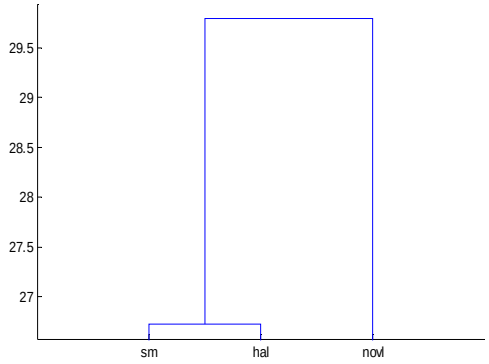

Figure 6. DSDistance clustering results on raw time series data (first cluster)

The clustering algorithm automatically divides these seven time series data into three clusters, one cluster has only four critical change-points (*EOG*, *CDIS* and *SCOX*), one has five critical change-points (*SM*, *HAL* and *NOVL*), and one has no change-point at all (*WMT*).

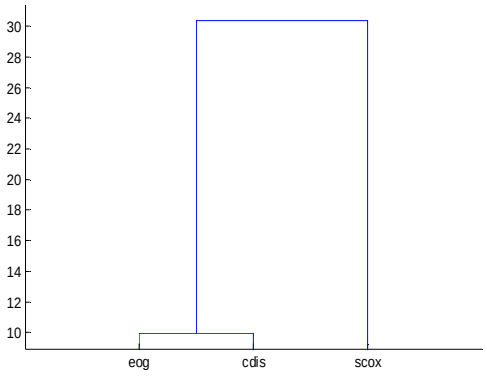

Figure 7. DSDistance clustering results on raw time series data (second cluster)

With the same compression ratio, PAA representation obtained the same clustering results as shown in fig. 7 and 8 with different threshold values when clustering EOG, CDIS and SCOX in one group and SM, HAL and NOVL in another group.

Fig. 2 and 3 show that Euclidean distance is sensitive to the different baseline and scale of time series, similar results also obtained by Shasha[2].

Based on Euclidean distance measure, the PAA representation of time series has the same clustering results on the raw time series data; A different clustering group on normalized time series based on PAA representation is obtained. The same compression ratio of PAA representation is the same as our critical change-points representation. Therefore, we can see that PAA representation does not lost the structure of time series on raw time series data, but not the normalized data when the comparable compression ratio is used to our data mapping method.

Our critical change-points representation achieves the same clustering results as the PAA mapping method under the same compression ratio. The approach we proposed adapts to the structure of time series automatically. The author of PAA approach also proposed an adaptive piecewise constant approximation[12].

The comparison of similarity based on raw data shows that the scaling and shifting have no affect on price movements comparison based on our distance measure that considers the relative position of corresponding change-points in the time series because PEuclidean shows the same results on normalized time series.

5 CONCLUSTIONS

In this paper, we introduced a new distance measure for clustering that considers the relative position of corresponding change-points in the time series. The distance measure proposed is suitable to deal with online similarity matching, such as data stream similarity matching, where traditional matching methods for time series are inefficient and the dimensionality reduction methods are very costly to apply repeatedly each time a new data arrives. Using our proposed method, it is also not necessary to keep statistics over the whole clustering process when new data come in. This distance measure is also not sensitive to the shifting and scale of the time series data.

We also proved that our distance measure is a general distance metric. It works as good as the Euclidean distance measure that uses normalized data, and the well defined PAA mapping approach. The performance closes to human perceptual judgment as well. The distance measure proposed is well suited to online time series data stream because it does not maintain stream statistics over data streams.

Future research can proceed to several directions. One is to use our distance measure only based on the landmarks[22] of the time series to reduce the computation time and dimensionality. The other is to incorporate indexing techniques in the searching algorithm because we have proved that our distance function is metric.

6 REFERENCES

[1] E. Keogh, K. Chakrabarti, M. Pazzani, and S. Mehoritra, "Dimensional Reduction for Fast Similarity Search in Large Time Series Databases," *Knowledge and Information Systems*, vol. 3, pp. 263-286, 2001 2001.

[2] D. E. Shasha and Y. Zhu, *High performance discovery in time series : techniques and case studies*. New York: Springer, 2004.

[3] R. Agrawal, C. Faloutsos, and A. Swami, "Efficient Similarity Search in Sequence Databases," *Proceedings of the 4th International Conference on Foundations of Data Organization and Algorithm, New York:Springer, 1993.*, 1993 1993.

[4] C. Faloutsos, M. Rangenathan, and Y. Manolopoulos, "Fast Subsequence Matching in Time-Series Database," in *Proc. ACM SIGMOD Conf.*, Minneapolis, 1994.

[5] Y.-S. Moon, K.-Y. Whang, and W.-S. Han, "General Match: A Subsequence Matching Method in Time-Series Databases Based on Generalized Windows," *SIGMOD*, pp. 382-393, 2002 2002.

[6] K.-p. Chan and A. W.-c. Fu, "Efficient Time Series Matching by Wavelets," *Proceedings of Internation Conference on Data Engineering (ICDE '99), Sydney*, p. 126, 1999.

[7] A. C. Gilbert, Y. Kotidis, S. Muthukrishnan, and M. J. Strauss, "Surfing Wavelets on Streams: One-Pass Summaries for Approximate Aggregate Queries," in *Proceedings of the 27th VLDB Conference*, Roma, Italy, 2001, pp. 79-88.

[8] Y. Huhtala, Kärkkäinen, J. & Toivonen, H. , " Mining for similarities in aligned time series using wavelets," *Data Mining and Knowledge Discovery: Theory, Tools, and Technology, SPIE Proceedings Series, Orlando, FL*, vol. 3695, pp. 150-160, Apr. 1999

[9] D. Wu, D. Agrawal, and A. E. Abbadi, "Efficient Retrieval for Browsing Large Image Databases," in *Proc. CIKM*, Rockville, MD., 1996, pp. 11-18

[10] F. Korn, H. V. Jagadish, and C. Falouts, "Efficient Supporting Ad Hoc Queries in Large Datasets of Time Sequences," in *SIGMOD*, 1997, pp. 289-300.

[11] E. Keogh and M. Pazzani, "Scaling up Dynamic Time Warping for Datamining applications," in *KDD* Boston, MA, 2000, pp. 285-289.

[12] E. Keogh, K. Chakrabarti, S. Mehoritra, and M. Pazzani, "Locally Adaptive Dimensionality Reduction for Indexing Large Time Series Databases," in *Proc. SIGMOD*, Santa Barbara, California, 2001, pp. 151-162.

[13] B.-K. Yi, H. V. Jagadish, and C. Falouts, "Efficient Retrieval of Similar Time Sequences under Time Warping," *ICDE*, pp. 201-208, 1998 1998.

[14] G. Das, D. Gunopulos, and H. Mannila, "Finding Similar Time Series," in *PKDD*, 1997, pp. 88-100.

[15] H. Wu, B. Salzberg, and G. C. Sharp, "Subsequence Matching on Structured Time Series Data," in *SIGMOD*, Baltimore, Maryland, USA, 2005.

[16] M. Datar, A. Gionis, P. Indyk, and R. Motwani, "Maintaining Stream Statistics over Sliding Windows," *SIAM Journal on Computing*, vol. 31, pp. 1794-1813, 2002.

[17] Y. Zhu and D. Shasha, "Efficient Elastic Burst Detection in Data Streams," in *SIGKDD* Washington, DC, USA: ACM, 2003.

[18] R. Jin and G. Agrawal, "Efficient Decision Tree Construction on Streaming Data," in *Conference on Knowledge Discovery in Data archive Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, Washington, D.C., 2003, pp. 571-576.

[19] S. Muthukrishnan, R. Shah, and J. S. Vitter, "Mining Deviants in Time Series Data Streams," in *Proceedings of the 16th International Conference on Scientific and Statistical Database Management (SSDBM 2004)*, Santorini Island, Greece, 2004.

[20] J. Gehrke, F. Korn, and D. Srivastava, "On Computing Correlated Aggregates Over Continual Data Streams," *SIGMOD*, pp. 126-133, 2001.

[21] J. Qu and H. R. Arabnia, "Mining Structural Changes in Financial Time Series with Gray System," in *DMIN 2005*, 2005.

[22] E. Keogh, "A Fast and Robust Method for Pattern Matching in Time Series Databases," in *Proceedings of 9th Internatinoal Conference on Tools and Artificial Intelligence (ICTAI)*, 1997, pp. 578-584.