THE INTERNATIONAL JOURNAL OF SCIENCE & TECHNOLEDGE

A Survey on Fast Pattern Matching in Stream Time Series Data for Electrical Applications

K. Manikandan
M. Tech. Student, Dr. Mgr Educational and Research Institute University, Tamil Nadu, India
Er. S. Bhuvaneswari
M.E., Assistant Professor, Dr. Mgr Educational and Research Institute University, Tamil Nadu, India

Abstract:
In real time application such as sweep frequency response analysis, Coal mine surveillance, Thermal power stations, multimedia data retrieval and privacy preserving data streams arrive at a rate higher than in traditional sensing applications. The processing of these raw data must be as fast as stream speed. These applications convert the raw data into a specified pattern and these patterns vary from application to application. It is then subjected to sophisticated query processing to extract high level information. Uncertainty in stream time series may occur for two reasons such as the inherent imprecision of sensor readings or privacy preserving conversion. Representation of uncertain data over stream time-series, uncertain data management and designing of a data mining algorithm on considering the uncertainty affects the data mining process. The processing of data in the real time application is difficult as it is naturally incomplete and noisy and the observed data pattern is different from the actual pattern required for further processing. A major challenge in processing the stream time-series data with uncertainty is to capture uncertainty as data propagates through query operators until the final result and to process the data at stream speed. The modeling of uncertain time-series without affecting the precision of the system still remains a difficult task. More importantly, it should avoid increased false positives. In this paper, a survey of time series data management for pattern matching is provided. The similarity search over uncertain data will be explored. The issues in existing research, limitations and methodology for pattern matching on uncertain time series stream data are examined in this paper.

Keywords: *Similarity pattern match, stream time series, multi scale segment median of Image approximation.*

1 Introduction

Stream time-series data management has become a hot research topic due to its wide range of applications such as sweep frequency response analysis, Coal mine surveillance, Thermal power stations (used in control and instrumentation) and multimedia data retrieval which require continuously monitoring stream time series

Compared to achieved data, stream time series have their own characteristic:

1. Data are frequently updated in stream time series.
2. Due to the frequent updates, it is very difficult to store all the data in memory or on disk, thus, data summarization and one pass algorithms are usually required to achieve a fast time response.

In this work we propose a novel approach to efficiently perform the pattern matching over stream time series. In order to save the computational cost and offer a fast response, we present a novel multi scale representation for time series, namely multi scale segment median of image (MSMI).

Most importantly, we propose a multi step filtering approach, with respect to the MSMI representation, to prune false candidates before computing the real distances between patterns and stream time series.

2. Related Work

This section overviews previous work on similarity search over archived time-series data and monitoring over stream time series.

In the research literature, many approaches have been proposed for the similarity search on the archived time series. The pioneering work by Agrawal [4] proposed the whole matching, which finds data sequences of the same length that are similar to a query sequence. Later, Faloutsos [5] extended this work to allow the subsequence matching, which finds subsequences in the archived time

series that are similar to a given query series. In these two works, Euclidean distance is used to measure the similarity between (sub) sequences. In order to perform an efficient similarity search, the GEMINI framework [5] is proposed to index time series and answer similarity queries without false dismissals. Since the dimensionality of time series is usually high (e.g., 1,024), the similarity search over high-dimensional index usually encounters a serious problem, known as the "curse of dimensionality." That is, the query performance of the similarity search over indexes degrades dramatically with the increasing dimensionality. In order to break such curse, various dimensionality reduction techniques have been proposed to reduce the dimensionality of time series before indexing them. But only DFT and DWT have been used in the scenario of stream time series, however, with the limitation that only one pattern is considered. Similarity measures. In addition to Euclidean distance (L2-norm)[4] [5] several other distance functions have been proposed to measure the similarity between two time series in different applications such as Dynamic Time Warping, longest Common Subsequence

, and Edit Distance with Real Penalty Specifically, Euclidean distance requires the time series to have the same length, which may restrict its applications. On the other hand, DTW can handle sequences with different lengths and local time shifting; however, it does not follow the triangle inequality, which is one of the most important properties of a metric distance function. A recent work makes DTW index able by approximating time series with bounding envelopes. The resulting R-tree index[6] is clearly inefficient for query processing on stream time series, in terms of both update and search cost. ERP can support local time shifting and is a metric distance function. LCSS[7] is proposed to handle noise in data; however, it ignores various gaps in between similar subsequences, which leads to inaccuracy. In contrast, our work focuses on Lp-norm, which covers a wide range of applications. Formally, the Lp-norm distance between two series X(X[0],X[1],….X[n-1]) and Y(Y[0],Y[1],…,Y[n-1]) of length n is defined as

$$Lp(X,Y)=\sqrt{\sum_{i=0}^{n-1}|X[i]-Y[i]^p}$$

where p>= 1. Note that L1-norm is also called Manhattan distance, whereas L2-norm is Euclidean distance.

2.1. Monitoring in Stream Time Series

Not much previous work has been published for monitoring stream time-series data. Zhu and Shasha[9] proposed a method to monitor the correlation among any pair of stream time series within a sliding window, in which DFT was used as a summary of the data. Later, they introduced a shift wavelet tree (SWT) based on DWT to monitor bursts over stream time-series data. Bulut and Singh[8] improved the technique by using multi scale DWT trees to represent data. Gao and Wang[3] proposed a prediction model to save the computational cost during the matching between a single stream time series and multiple static patterns. Recently, Papadimitriou[10] proposed a method for capturing correlations among multiple stream time-series data with the help of an incremental PCA computation method. With respect to data estimation, estimated the current values of a co-evolving time series through a multivariate linear regression. These approaches, however, are different from our similarity match problem, in the sense that they do not assume any patterns available. Instead, they aim at detecting either patterns in a stream time series or changes in correlation pattern over multiple stream time series. Moreover, Wu[11] proposed an online matching algorithm for detecting subsequences of financial data over a dynamic database. However, their segmentation and pruning methods are designed for financial data only and cannot be applied to detect general patterns in stream time series.

3. Issues in the Existing Research Work

Most of the time-series representation methods do not reduce the dimensionality of the original data. The distance measures defined in the symbolic representation method have little correlation with that of in the original time-series. Among the dimensionality reduction techniques mentioned above, only DWT and DFT deals with the stream time-series. The existing similarity measures do not support online monitoring applications as these are designed for full sequence matching. The existing pattern matching algorithms are susceptible to noise, offset translation, and amplitude scaling with various degrees.

The existing approaches for modeling uncertain time-series are based on only two different approaches. The first approach estimates the probability density function over uncertain values based on the prior knowledge. The second approach summarizes the distribution of uncertain data as a result of frequent measurements. These approaches offer results with increased false positives. The real challenge is achieving tradeoff between accuracy and efficiency in uncertain time series data streams.

3.1. Limitations

The previous works on pattern matching on archived time series are not suitable for stream time series as they cannot deal with the frequent updates. Concerned with the data streams, the efficiency of the computation process is the key issue as the data can be processed only once during the entire computation in the presence of uncertain data. Generally, processing of raw uncertain data is difficult because of its mass volume and stringent time schedule. Raw uncertain data generally comprise of noise that interrupts the pattern matching process and produces false positives. All the time series dependent applications require immediate responses and cannot prefer any post processing. Striking solution between the accuracy and efficiency is highly challenging in the presence of uncertainty.

4. Methodology

The design and development of a pattern matching system of the data stream that captures data uncertainty from data collection to query processing to obtain final pattern that matches the query pattern. The incoming patterns are initially transformed into streams with uncertain data to model the core data generation process. The data of interest can be inferred from this model. The inference process computes a distribution of values for the uncertain data required in later processing. The uncertain data are then eliminated to prune off large search space using any of the lower bound distance measures. The size of the search space varies exponentially with the number of features in the query pattern. Therefore, it becomes necessary to depend on effective search techniques. The sensor data are generally updated in a periodic manner. This necessitates the use of indexing system to update the current attribute value. Each part of the query is associated with a range of possible values and probability density function, which quantifies the behavior of the data over that range. The obtained resultant pattern is compared with the time-series database to provide the final pattern that matches with the query pattern. Table 1 shows the previous works on time series data.

5. Conclusion

The area of uncertain data management in stream time series application has been treated as an unsolved issue. This work presented an overview of pattern matching methods over uncertain time series stream data. This work also presented the commonly used similarity measures, similarity searches on stream time series and uncertain data along with the representational issues in uncertain data management. The accuracy of the pattern matching algorithm is measured in terms of precision and recall. If the uncertainty is perfectly mined out, stream time series data can be extended to numerous applications

6. References

i. Aggarwal. C, 2008. On Unifying Privacy and Uncertain Data Models. IEEE 24th international Conference on data engineering, pp: 386– 395.

ii. Agrawal. R, C. Faloutsos, and A.N. Swami, 1993. Efficient Similarity Search in Sequence Databases. Proceedings of 4th International Conference on Foundations of Data Organization and Algorithms (FODO).

iii. Anthony Bagnall, Chotirat Ann Ratanamahatana, Eamonn Keogh, Stefano Lonardi, and Gareth Janacek, 2006. A Bit Level Representation for Time Series Data Mining with Shape Based Similarity. Springer, Data Mining and Knowledge Discovery.

iv. Berndt. D.J and J. Clifford, 1996. Finding Patterns in Time Series: A Dynamic Programming Approach. Advances in Knowledge Discovery and Data Mining.

v. Boreczky. J.S and L.A. Rowe, 1996. Comparison of Video Shot Boundary Detection Techniques. Proceedings of 18th international Symposium on Storage and Retrieval for Image and Video Databases.

vi. Cai. Y and R. Ng, 2004. Indexing Spatio-Temporal Trajectories with Chebyshev Polynomials. Proc. ACM SIGMOD.

vii. Chan FKP, AWC Fu, C Yu, 2003. Haar wavelets for efficient similarity search of time-series: with and without time warping. Knowledge and Data Engineering, IEEE Transactions on Volume: 15,Issue:3,pp: 686 – 705.

viii. Chan. K.P and A.W.-C Fu, 1999. Efficient Time Series Matching by Wavelets. Proceedings of 15th International Conference on Data Engineering (ICDE).

ix. Charu C. Aggarwal and Philip S. Yu, 2009. A Survey of Uncertain Data Algorithms and Applications. IEEE transactions on knowledge and data engineering, vol. 21, no. 5.

x. Charu C. Aggarwal, Philip S. Yu, 2008. A Framework for Clustering Uncertain Data Streams. IEEE 24th international conference on data engineering, pp: 150- 159.

xi. Chen. L and R. Ng, 2004. On the Marriage of Edit Distance and Lp Norms. Proc. 30th Int'l Conf. Very Large Data Bases (VLDB).

xii. Chen. Q, L. Chen, X. Lian, Y. Liu, and J.X. Yu, 2007. Indexable PLA for Efficient Similarity Search. Proc. 33rd Int'l Conf. Very Large Data Bases (VLDB).

xiii. Eamonn Keogh and Padhraic Smyth, 1997. A Probabilistic Approach to Fast Pattern Matching in Time Series Databases. Proceedings of the 3rd international conference of Knowledge Discovery and Data Mining, pp: 20- 24.

xiv. Eamonn Keogh Kaushik Chakrabarti Michael Pazzani Sharad Mehrotra, 2001.Dimensionality Reduction for Fast Similarity Search in Large Time Series Databases. Springer.

xv. Eamonn Keogh, 1997. A Fast and Robust Method for Pattern Matching in Time Series Databases.

xvi. Faloutsos. C, M. Ranganathan, and Y. Manolopoulos, 1994. Fast Subsequence Matching in Time-Series Databases. Proc. ACM SIGMOD.

xvii. Graham Cormode, and Andrew McGregor, 2008. Approximation Algorithms for Clustering Uncertain Data. Proceedings of the 27th ACM SIGMOD-SIGACT-SIGART symposium on principles of database system, pp: 191- 200.

xviii. Gullo F, G Ponti, A Tagarelli, S Greco, 2009. A time series representation model for accurate and fast similarity detection. Pattern Recognition, Elsevier Volume 42, Issue 11, Pages: 2998–3014.

xix. Guttman. A, 1984. R-Trees: A Dynamic Index Structure for Spatial Searching. Proc. ACM SIGMOD.

xx. Jessica Lin, Eamonn Keogh, Li Wei, and Stefano Lonardi, 2007. Experiencing SAX: a Novel Symbolic Representation of Time Series. ACM journal on Data Mining and Knowledge Discovery, Vol. 15, Issue 2, pp: 107- 144.

xxi. Johannes Abfalg, Hans-Peter Kriegel, Peer Kroger, Matthias Renz, 2009. Probabilistic Similarity Search for Uncertain Time Series. IEEE international conference on information engineering and computer science, pp: 1-4,

xxii. Karen Zita Haigh, Wendy Foslien, Valerie Guralnik, 2002. Visual Query Language: Finding patterns in and relationships among time series data. Tech Report.

xxiii. Keogh. E, 2002. Exact Indexing of Dynamic Time Warping. Proc. 28th Int'l Conf. Very Large Data Bases (VLDB).

xxiv. Kiyoung Yang and Cyrus Shahabi, 2004. A PCA-based similarity measure for multivariate time series. MMDB '04 Proceedings of the 2nd ACM international workshop on Multimedia databases, Pages: 65 – 74.

xxv. Korn. F, H. Jagadish, and C. Faloutsos, 1997. Efficiently Supporting Ad Hoc Queries in Large Datasets of Time Sequences. Proc. ACM SIGMOD.

xxvi. Li Q, B Moon, IFV Lopez, 2004. Skyline index for time series data. Knowledge and Data Engineering, IEEE Transactions on (Volume: 16, Issue: 6), Pages: 669 – 684.

xxvii. Megalooikonomo. V, Q. Wang, G. Li, and C. Faloutsos, 2005. A Multiresolution Symbolic Representation of Time Series. Proc. 21st Int'l Conf. Data Eng. (ICDE).

xxviii. Michele Dallachiesa, Besmira Nushi, Katsiaryna Mirylenka, 2011. Similarity Matching for Uncertain Time Series: Analytical and Experimental Comparison. Proceedings of the 2nd ACM SIGSPATIAL international workshop on querying and mining uncertain spatio-temporal data, pp: 8-15.

xxix. Qiang Wang and Vasileios Megalooikonomou, 2008. A Dimensionality Reduction Technique for Efficient Time Series Similarity Analysis. Information Systems Volume 33, Issue 1, Pages: 115–132.

xxx. Sarangi. S and K. Murthy, 2010. DUST: a generalized notion of similarity between uncertain time series. Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, pages: 383–392.

xxxi. Sethukkarasi, R. ; Rajalakshmi, D. ; Kannan, A, 2010. Efficient and Fast Pattern Matching in Stream Time Series Image Data. ICIIC IEEE conference publications, Page(s): 130 – 135.

xxxii. Spiros Papadimitriou, Jimeng Sun, and Christos Faloutsos, 2005. Streaming Pattern Discovery in Multiple Time-Series. ACM proceedings of the 31st international conference on very large databases, pp: 697- 708.

xxxiii. Suresh RM, K Dinakaran, P Valarmathie, 2008. Clustering Gene Expression Data Using Self- Organizing Maps. Journal of Computer Applications, Vol. 1, no. 4, Pages: 5-7.

xxxiv. Tomoya saito, takuya kida, and hiiroki arimura, 2007. An efficient algorithm for complex pattern matching over continuous data streams based on bit-parallel method. IEEE international workshop on databases for next generation researchers, pp : 13-18.

xxxv. Valarmathie P, Dr.M V Srinath, K Dinakaran, 2009. An increased performance of clustering high Dimensional data through dimensionality Reduction technique. Journal of Theoretical and Applied Information Technology.

xxxvi. Xiang Lian, Lei Chen, Jeffrey Xu Yu, Jinsong Han and Jian Ma, 2009. Multiscale Representations for Fast Pattern Matching in Stream Time Series. IEEE transactions on knowledge and data engineering, vol. 21, no. 4.

xxxvii. Xiaoyue Wang, Abdullah Mueen, Hui Ding, Goce Trajcevski, Peter Scheuermann, and Eamonn Keogh, 2013. Experimental comparison of representation methods and distance measures for time series data. ACM journal on Data Mining and Knowledge Discovery, Vol. 26 , Issue 2, pp: 275- 309.

xxxviii. Xue. W, Q. Luo, L. Chen, and Y. Liu, 2006. Contour Map Matching for Event Detection in Sensor Networks. Proceedings of ACM SIGMOD.

xxxix. Yanlei Diao, Boduo Li, Anna Liu, Liping Peng, and Charles Sutton, 2009. Capturing Data Uncertainty in High Volume Stream Processing.

xl. Yeh. M, K. Wu, P. Yu, and M. Chen, 2009. PROUD: a probabilistic approach to processing similarity queries over uncertain data streams. Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology, pages: 684–695, ACM.

xli. Zhao. Y, C. C. Aggarwal, and P. S. Yu, 2010. On wavelet decomposition of uncertain time series data sets. CIKM, pages: 129–138.

xlii. Zhu. Y and D. Shasha, 2003. Efficient Elastic Burst Detection in Data Streams. Proc. Ninth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD).

xliii. Zhu. Y and D. Shasha, 2003. Warping Indexes with Envelope Transforms for Query by Humming. Proc. ACM SIGMOD.