Study on Visual Techniques of Potential Pattern Discovery for Time Series Data

Dalin Xu[1] and Yingmei Wei[1,a]

[1]*College of Systems Engineering, National University of Defense Technology,410073 Changsha, China*

Abstract. Sequential pattern mining is always a very important branch of time series data mining. The pattern mining with visual means can be used to extract the knowledge of time series data more intuitively. Based on the research content, this paper analyzes the sequence pattern mining methods in different aspects and their combination with visualization technology. We further discuss and summarize the advantages of different visualization methods in discovering the potential patterns in time series data. Different systems and models have their unique information to show the focus. Compared with the characteristics of the model, the development and evolution of visualization technology for the discovery of potential patterns of time series data can be summarized. Finally, this paper discusses its development trend and how to play a greater role in the era of big data.

1 Introduction

The study of the time series data mining consists mainly of feature representation, similarity method, and query, sequence pattern mining, clustering and classification, anomaly detection, prediction, etc. Therein, feature representation, similarity method, and sequential pattern mining are the key and hard problems in the study of time series. Because the time sequence is typically a high-dimensional data, it's necessary to reduce the amount of calculation by using the lower dimensions to gain global or local characteristics. The similarity measures are used to calculate the distance between the two sequences, the similarities between the sequences, and then provide the distance based on other data mining algorithms. While sequential pattern mining can effectively find the hidden connections in multiple sequences, and mining frequent sequences as patterns can be used for temporal association rule discovery or fault sequence detection.

The task of sequence pattern mining is to find out that there exist regular pattern sequences in a given database. Sequence pattern mining is widely used in business management, production control, market analysis, engineering project design, and scientific exploration areas, which include purchase behavior pattern analysis, gene sequence analysis, web access pattern prediction, and so on. Sequence pattern mining and association rule mining have a lot of similarities. However, the research emphasis is different, association rule mining is concerned with the same transaction data of the correlation between multiple projects in question, and sequence pattern mining is more concerned with the relation between different transaction data sequence. From traditional frequent pattern mining based on extended research and development, the sequence pattern mining main research contents can be divided into closed pattern mining, cycle pattern mining and motif pattern mining these a few respects [1]. This article will analyze on these aspects of pattern mining technology, and combined with the visualization technology to further explore the potential model time-series data found that the evolution regularity of visualization technology and its development trend in the era of big data. However, due to the application background of the sequential data visual analysis is extensive, at the same time sequence data has the complexity of timing data, such as temporal variability, high dimension, noise disturbance and fluctuation, so time-series data visual analysis has also been a hot and difficult problem in the field of information visualization research.

2 Traditional frequent pattern mining

Traditional pattern mining, which is to find all the frequent subsequences in the given sequence database, is a traditional method of mining based on the support model. Inchoate sequential pattern mining algorithms are mostly based on a classical association rule mining algorithm based on Apriori proposed to the frequent mode of any sub-mode are frequent prior knowledge, generate frequent iteration k+1 frequent k sequence by sequence, thereby generating all frequent modes. According to this enlightenment, researchers suggested a series of class Apriori algorithms, such as Apriori All, Apriori Some, and Dynamic Some. Wijk et al. provided a visual method with which combining the Aprioi algorithm [2]. In other word, due to a kind of visualization combined with calendar after cluster

[a]Corresponding author: weiyingmei126@126.com

analysis, the univariate time series data can be deeply mined. This work shows the daily patterns as graphs, moreover, displaying the class clusters on the calendar. In this way, a combination representation of daily patterns and cluster clusters is created. Fig. 1 shows that the results of time-series data on the number of employees in the ECN cluster analysis. The figure shows that the most significant seven clusters, wherein each color chart on the right represents the average value of the cluster, the left side portion thereof colored according to the cluster for each day of the calendar belongs. In terms of visual interaction, the users can manually or automatically research the relationships between variables by extending interactive visualization and analyzing several variables at the same time.

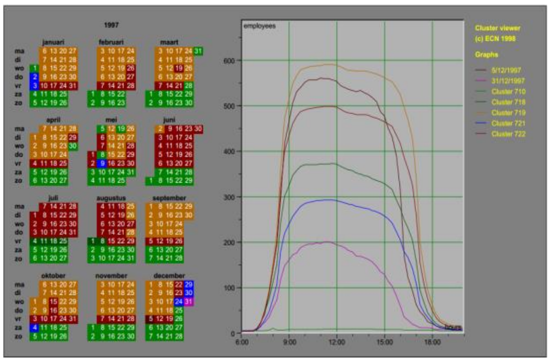
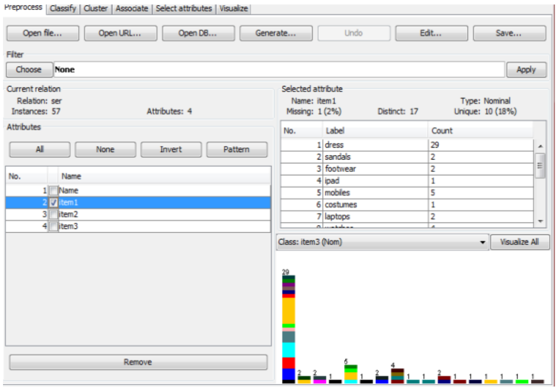


Figure 1 Calendar view of the numbers of employees

Swamy et al. [3] use the Apriori algorithm and Weka tools to summarize various data mine techniques for frequent project search patterns, such as association classification, minconf and candidate generation. Due to this, it further combines the visual means to analyze the frequent project search patterns. This paper's application background is on the market basket of business processes in the frequent mode identification and analysis of the project, while the market basket analysis is generally used for this type of business improvement and introduction of new products, research by comparing different transactions to determine how different decision-making transactions and how business needs change, that is, for the most relevant projects are identified. This article uses the Apriori algorithm to generate new list items and alternate keys, depending on the confidence factor, and further uses mining Weka tools and visualizations to observe different sets of items and identify the most frequent search items, as shown in Fig.2.



Filter: None

Current relation
Relation: ser
Instances: 57 Attributes: 4

Attributes

No.		Name
1	□	Name
2	☑	item1
3	□	item2
4	□	item3

Selected attribute
Name: item1 Type: Nominal
Missing: 1 (2%) Distinct: 17 Unique: 10 (18%)

No.	Label	Count
1	dress	29
2	sandals	2
3	footwear	2
4	ipad	1
5	mobiles	5
6	costumes	1
7	laptops	2

Class: item3 (Nom)

Figure 2 Analysis results of using weka tools

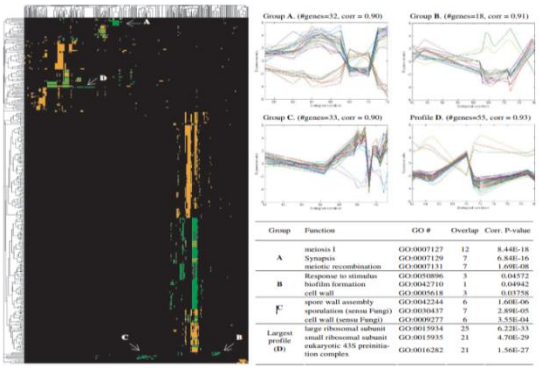
Although the traditional methods of frequent pattern mining have been studied extensively, there are still some shortcomings. Firstly, the time series containing a large amount of noise, and the part of the original values should be the same point. However, due to the influence of noise environment, resulting in a different degree of fluctuation, and causing the inconsistent values. In the case of statistical support, they cannot be added to the count; if support is less than a certain threshold, they cannot be output as frequent sequences. As a result, longer frequent sequences are difficult to unearth. Moreover, when the data volume is large, or the support threshold is small, the frequent sequence number will increase exponentially, and the performance of the algorithm decreases sharply. In order to satisfy more application demands, the specific pattern mining methods, such as closed patterns, periodic patterns, and motif patterns, are more and more concerned by people.

3 Closed pattern mining

Closed pattern mining is an improvement for the traditional pattern mining. Only the maximum frequent sequences with the same degree of support are excavated, and the result set is compressed on the premise that the information is lossless. The typical closed pattern mining algorithm is a modified algorithm based on the Prefix Span algorithm by Yan et al. and Clo Span [4]. The algorithm will be able to generate a sequence of closed sequences to be stored in the result tree of the Hash index, and use the common prefix, the tracer pattern and the supermodel to enhance the algorithm's efficiency.

Gyenesei et al. [5] proposed a method of Mining Attribute Profile (MAP). It is a new similarity measure which can be applied together with the hierarchical clustering method on the basis of the Association Pattern Discovery (APD) discovery method, and this led to the discovery of hidden patterns of co-regulatory genes that traditional APD methods could not find. In this article, two well-known yeast microarray data sets are used to test the effectiveness of the method. The left half of Fig. 3 shows the cluster results of the

Yeast80 dataset, and in the right half shows the selected hidden gene spectrum and its biological relevance. We can find that belongs to the group A schema contains many involved in homologous chromosomes (synapses), and other events genes during meiosis. While group B and even group C much more strongly suggests that these patterns contain genes involved in the cell wall formation, and it largely because these patterns in biology-related information through altogether regulate gene significantly enriched with similar functions to prove. The experimental results show that MAP can effectively analyze gene expression data and can compete with double clustering technology.



Group	Function	GO #	Overlap	Corr. P-value
A	meiosis I	GO:0007127	12	8.44E-18
	Synapsis	GO:0007129	7	6.84E-16
	meiotic recombination	GO:0007131	7	1.69E-08
B	Response to stimulus	GO:0050896	3	0.04572
	biofilm formation	GO:0042710	1	0.04942
	cell wall	GO:0005618	3	0.03758
F	spore wall assembly	GO:0042244	6	1.60E-06
	sporulation (sensu Fungi)	GO:0030437	7	2.89E-05
	cell wall (sensu Fungi)	GO:0009277	6	3.55E-04
Largest profile (D)	large ribosomal subunit	GO:0015934	25	6.22E-33
	small ribosomal subunit	GO:0015935	21	4.70E-29
	eukaryotic 43S preinitiation complex	GO:0016282	21	1.56E-27

Figure 3 MAP of genetic data

Kiraly et al. [6] developed a new FCPMiner pattern mining method. This new method is unique in frequent closed itemsets in the mining of the binary data scalability, which further combining visualization method and model aggregation method for the most meaningful, non overlapping mode for testing. To provide a visualization technique for the thousands of scattered subsets of the original data, a novel technique for visualizing the original data matrix by reordering the rows and columns based on the discovered closed patterns has been presented, and to further achieve the goal of visualization raw data matrix. Visual results as shown in Fig. 4, the distance changes through the grey-scale value after the rearrangement to the similarity between the said model, the deeper the color represents, the similarity is higher. As you can see from the results, this paper puts forward the polymerization can be effectively for has a strong correlation between mining frequent closed patterns.

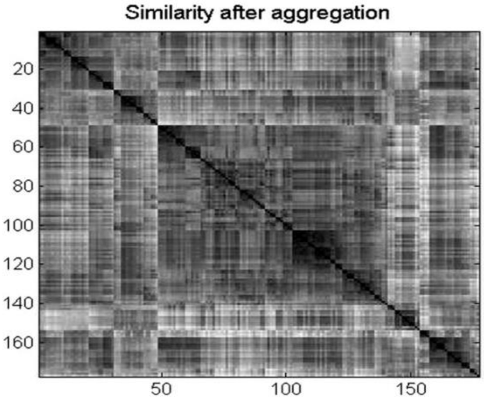


Figure 4 Visual results reflecting FCPMiner's efficiency

4 Periodic pattern mining

Periodic pattern mining is a category of evolution analysis, which is an effective sequence mode mining method for extracting and analyzing sequence patterns in time series with periodic behavior. So far, the full cycle analysis and partial periodic analysis of the two main studies that have been excavated in the cycle pattern. Therein, the full-cycle analysis mining is mainly applied to periodic detection, Fast Fourier Transform (FFT) and other approaches, and has been studied extensively in statistics and signal analysis. The partial cycle analysis mining is based on Apriori and constrained mining algorithm to carry out corresponding improvements.

Chen et al. [7] designed a Spatio-Temporal Visualization (STV) system which is for helping identify criminal patterns. This system integrates three visualization technologies for the synchronous views, and they are: GIS view, timeline view, and periodic pattern view. In this case, the major aim of the periodic pattern view is to offer a quick and easy way for the CIO to search for the time-crime pattern, and it has three sub-tools: the periodic pattern tool, the histogram tool, and the line chart tool. In periodic mode tool, the event by the user to select the granularity of the summary and crime types, and the circle is used to represent the time granularity, within the circle of the sector is the period of time. For each department, the event is summarized in the form of a crime and it shows the peak, and the length of the peak is the relative quantity of the event -- and the benefit of doing that is that the analyst can see that there are different patterns at different times. The histogram tool and the fold chart tool are used as a conventional chart tool and may show the distribution of criminal events over time.

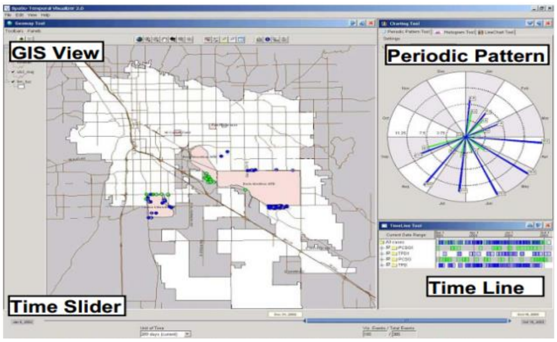


Figure 5 Visual analytic system of STV

Siripatana et al. [8] suggested the "3D Spring Model" to visualize multidimensional time series data related to the weather direction. The aim is to discover patterns behind large time series weather data sets and to clear seasonal structures in the data. The model through a direct comparison between the continuous spring supports seasonal change and the anomaly of visibility, which applying the plan of Level-Of-Detail, which can be adjusted, to provide users with different pattern time focus. As shown in Fig. 6 shows the Nakhon Si Thammarat in late March 2011 flood events of the case study, starting from the flooding events of 21/03, the wind, temperature, and rainfall three parameters have obvious decline. Spring model is highly self-contained accumulate data for a long time. It has that capability of interactivity, flexibility and user friendliness. It is very suitable for the visualization environment of high computational power, and also has a wide application capability.

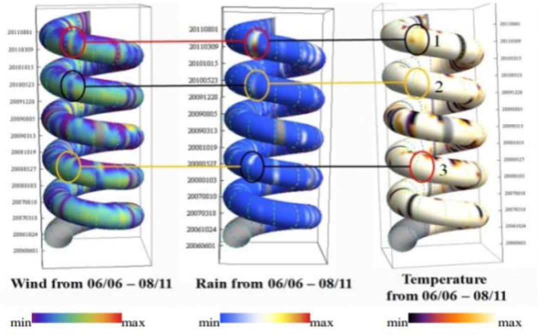


Figure 6 The display of timing data from 3D Spring Model

5 Motif pattern mining

From its nature, both closed pattern mining and periodic pattern mining are developed on the basis of frequent pattern mining. The elements of frequent sub-sequences mined can be discontinuous, and there just is a sequential time sequence. While the motif pattern mining is different from the former two, its goal is to discover similar two or more consecutive sub-sequences in the time series, which involves the similarity measurement between sequences, and it can deal with the noise in the sequence and the deformation (such as the situation of compression stretching). The concept of motif was first proposed by Patel et al. in 2002, suggesting that motif is a typical disjoint subsequence [9].The motif pattern mining originates from the analysis of biological gene sequence, which is widely concerned by experts and scholars both at home and abroad because of its feature representation, similarity measurement and index mechanism.

Li et al. [10] developed a motif visualization system which on the basis of grammar induction, as shown in Fig. 7. It can effectively recognize repeated patterns without knowing the length of patterns in advance. The upper-left image shows the input time series, and the upper-right corner is “control panel”, allowing the user to enter various options such as Sax parameters, initial window size, and the options of sliding window. The tool has the ability to discover hierarchies, regularities, and grammars from data, and also allows users to navigate and explore variable lengths of changeable lengths that coexist in the data set. Experimental results show that the grammar-based approach can find some important topics. Moreover, based on this, a search heuristic method is proposed to improve the quality of the induced grammar.

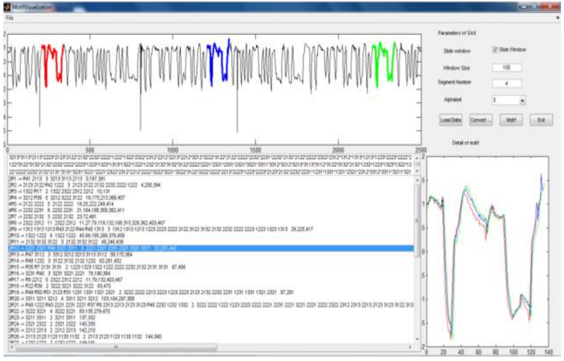


Figure 7 The syntax/motif visualization system based on the winding data set

Hao et al. [11] provided a visualization method for motif pattern mining. At first, they introduced a new motif discovery algorithm based on cluster analysis, event coding, frequent motif mining and efficiency characterization of these themes. What’s more, four new visualization methods have been introduced: motif layout - using the appearance of visual pattern of colored rectangles and hierarchical relationship; motif twist - enlarging or shrinking motifs for visualizing them more clearly; motif merge - combining multiple instances of the same adjacent pattern to simplify the display; pattern preserving prediction - using a pattern-preserving smoothing and prediction algorithm to provide a reliable prediction for seasonal data. This way allowing users adjust the distortion and merge to generate the best view on a single display. Also, by applying real-world data set of experimental results show that service managers are able to cross-examine topics and gain new insights into repetitive patterns to analyze system operations.

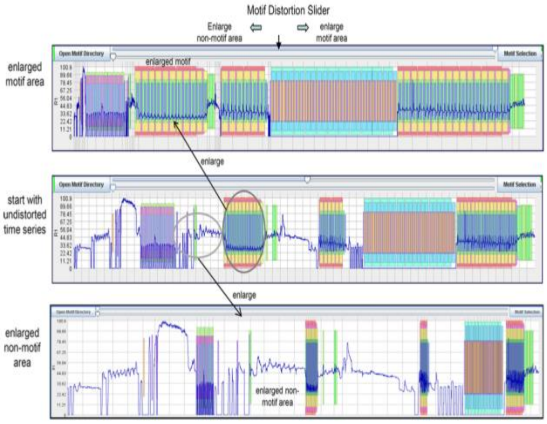


Figure 8 Interactive analysis of motif

6 Pattern mining of other kinds

Except the above-mentioned typical pattern mining methods, the scholars have combined the visualization methods with the anomaly mode monitoring, the community behavior pattern extraction and the event sequence pattern mining and so on. Through the verification of the algorithm by visual means has achieved multifaceted and multi-level demonstration effect.

Cao et al. [12] proposed a visual interaction system and framework - Voila (visual analysis of spatiotemporal data) , which was presented to interactively detect the anomaly patterns of spatiotemporal data collected from stream data sources. The system as shown in Fig. 9 is based on a new tensor-based anomaly analysis algorithm. Firstly, this algorithm transforms the spatiotemporal flow data into tensor time series, and analyzes the pattern based on the historical data, then achieves the detection of the anomaly in the online region of the context through the context analysis on the basis of the tensor decomposition. Taking smart cities as an example, the system has visual and interactive design capabilities that dynamically generate contextual, explanatory data summaries and allow for an interactive arrangement of exception patterns based on users' input.

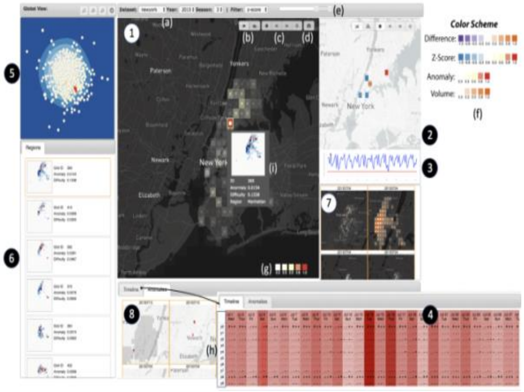


Figure 9 The interaction interface of Volia system

Wu et al. [13] presented an interactive visual analysis system, TelCoVis. It can help analysts use their domain knowledge to absorb the enlightenment of the simultaneous occurrence of urban population flows based on Power Grid's data. Meanwhile, this is the research and visual analysis of potential patterns in sequential data. This system combines various views, for example, the contour-based tree, the parallel coordinate system, the matrix graph and the extended line-up-type chart, which can be explored from many aspects in the common occurrence pattern exploration of urban population flow. The outline-based tree is shown in Fig. 10, which is used as an example of how many people visit the place at different times of the day. The distance between the region and the region where they from, and the "loyalty" of these areas (loyalty here means the frequency of access of people from a particular region) are characterized by the spatial and temporal characteristics of human movement in a given place. In this figure, (a), (b), (c), (d) respectively represent the time characteristics of population movements on office buildings, the shopping center, residential areas and the MTR station.

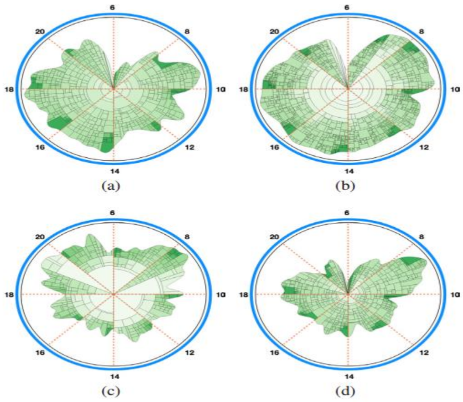


Figure 10 Contour-based tree

As shown in Fig.11, Chen et al. [14] presented a multi-layered visual analysis framework for event sequence data, which offers an exploration of interactive data. Firstly, a theory of information method on the basis of the Minimum Description Length principle (MDL) is presented to construct a rough summarize of event sequence data and to balance the information loss. This mean allows both sequence clustering and pattern extraction to occur simultaneously, and highly tolerates noise such as missing or additional events in the data. This is further proof that it supports the soft model, and it can contain different editing operations. Also, it's a case study of the availability and validity of the empirical approach by using two real-world data sets.

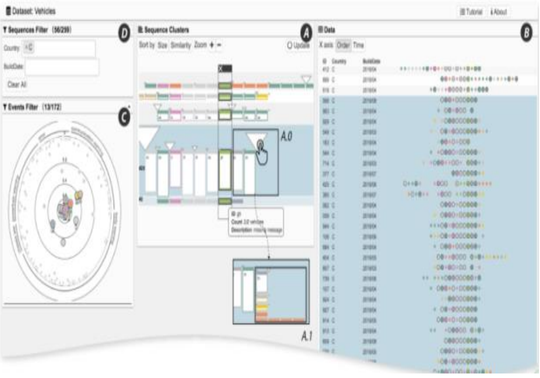


Figure 11 The interactive visualization interface of event sequence data pattern mining

7 Summary

When the previous visualization techniques are used in data mining, they are more used as an expression tool to generate views, while the analysis approach itself doesn't include visualization. Data visualization technology has the characteristics of flexibility and interactivity, and it contains all stages of the life cycle of data mining analysis, which are data preparation, model generation, knowledge usage, information visualization, and visual analysis as the end. These enable users to explore the data interactively, to participate in the analysis process, and to find useful patterns and information from big data volumes easier and more intuitive. As a result, visualization technology has increasingly become an integral part of the data analytics landscape.

Recently, the integration of knowledge discovery and visualization technology have aroused the attention of researchers, more and more scholars have started researches in this field. Based on the differences in the research and the focus, this article has analyzed and presented a series of the sequence pattern excavation methods and the integration of visual technology from different aspects. This paper has also discussed the advantages of various visualization methods in pattern mining in different fields, and introduced the methods that can be combined with other models to solve the problems related to the explosion of multi-dimensional time-varying information and different user demands.

References

1. Mooney, Carl H., and J. F. Roddick. "Sequential pattern mining -- approaches and algorithms." *Acm Computing Surveys* (2013).
2. Wijk, Jarke J. Van, and E. R. V. Selow. "Cluster and Calendar Based Visualization of Time Series Data." *IEEE Symposium on Information Visualization IEEE* (2002).
3. K. Swamy, G. Babu, R. Venkatasubbaih,"Identification of Frequent Item Search Patterns Using Apriori Algorithm and Weka Tool", *International Journal of Innovative Technology and Research*:2401-2403 (2015).
4. Yan, X. "CloSpan : Mining closed sequential patterns in large datasets." *Siam International Conference on Data Mining*:166-177 (2003).
5. Gyenesei, Attila, et al. "Mining co-regulated gene profiles for the detection of functional associations in gene expression data." *Bioinformatics*:1927-1935 (2007).
6. Király, András, et al. "Novel techniques and an efficient algorithm for closed pattern mining." *Expert Systems with Applications*:5105-5114 (2014).
7. Chen, Hsinchun, et al. "Visualization in law enforcement." *CHI '05 Extended Abstracts on Human Factors in Computing Systems ACM*:1268-1271 (2005).
8. Siripatana, Adil, et al. "The development of interactive 3D spring visualization for periodic multidimensional direction time-series data sets." *International Conference on Electrical Engineering/electronics, Computer, Telecommunications and Information Technology IEEE*:1-4 (2012).
9. Patel, Pranav, et al. "Mining Motifs in Massive Time Series Databases." *IEEE International Conference on Data Mining* (2002).
10. Yuan Li, Jessica Lin, and Tim Oates. "Visualizing variable-length time series motifs." *Proceedings of the 2012 SIAM International Conference on Data Mining*:895-906 (2012).
11. Hao, Ming C., et al. "Visual exploration of frequent patterns in multivariate time series." *Information Visualization*:71-83 (2012).
12. Cao, N., et al. "Voila: Visual Anomaly Detection and Monitoring with Streaming Spatiotemporal Data. " *IEEE Transactions on Visualization & Computer Graphics*:23-33 (2017).
13. Wu, W., et al. "TelCoVis: Visual Exploration of Co-occurrence in Urban Human Mobility Based on Telco Data. " *IEEE Transactions on Visualization & Computer Graphics*:935-944 (2015).
14. Chen, Y., P. Xu, and L. Ren. "Sequence Synopsis: Optimize Visual Summary of Temporal Event Data.

" *IEEE Transactions on Visualization & Computer Graphics* :1-1 (2017).