Membrane Technology
ISSN (online): 1873-4049

Trend Events and Lag Coefficients in Time Series Association Rule Discovery

Zhonglin Sheng*

Institute of Minerals Research, University of Science & Technology Beijing, Beijing, China.
**Corresponding Author.*

Abstract:
Trend events reveal long-term changes in time series data and help capture long-term pattern shifts. Lag coefficients, on the other hand, focus on the impact of past events on the present or future, helping to uncover time delays and causal relationships. In time series association rule discovery, considering trend events and lag coefficients can enhance the model's predictive power and provide more accurate decision support, particularly in fields such as finance, market analysis, and epidemiology. Both of these factors are key elements for effective modelling and rule mining, and are crucial for extracting meaningful time series patterns. This paper analyses those key challenges and introduces a new approach to solve these problems by constructing trend events and solving for lag coefficients. Based on this approach, some algorithms are proposed to improve classic association rule discovery method. The experiment with PM2.5 data set proves the effects of discretization of trend event and shows the hidden frequency pattern considered lag coefficient. This research represents a meaningful attempt to improve the effect of association rules discovery within multiply time series.

Keywords: Time Sequence Discretization, Association Rules Discovery, Trend Event, Lag Coefficient, Hidden Frequent Patterns

INTRODUCTION

Trend events typically refer to the occurrence of specific patterns or trends in a time series. They reflect long-term upward or downward trends in the data, or periodic fluctuations. These events are crucial for mining association rules in time series, as they can reveal key patterns of change and help predict future trends. The association rule discovery study, as an important area of data mining, usually focus on exploring the implicit and interesting relationships among transactions. Time series association rule discovery methods are primarily used to identify meaningful, time-ordered associations present in time series data. By analysing data at different time points, researchers can uncover potential patterns and trends, which can then be applied in decision-making. The research on pattern recognition in time series data began with Agrawal's introduction of the Generalized Sequential Pattern (GSP) mining method [1]. Related research has been ongoing, with Han et al. proposing the Free Span algorithm in 2000 [2], Pei introducing the Prefix Span algorithm in 2001 [3], which outperforms GSP and mining sequential patterns [4], and Zaki presenting the SPADE algorithm in the same year [5]. Harms and others proposed the MOWCATL algorithm for frequent pattern mining in multiple sequences, marking the first attempt to address the lag phenomenon encountered when discovering frequent patterns in multiple sequence data [6]. Rad and Sun identified lag by defining cross-correlation functions, while Yang Ren et al. [7] used the nonlinear least squares method to identify higher-order systems with delays. In 2016, Cai and others proposed a multi-time series association rule mining algorithm [8] that considers the distribution characteristics of the consequent. It is evident that after many years of research, association rule discovery algorithms have gradually expanded to cover time series data, but they have not yet achieved the same level of breadth and depth as traditional transactional data.

Compared to other transactional data, time series data has typical characteristics, the most important of which is the clear sequential relationship between all data nodes. According to the Markov principle, in a single entity sequence, the preceding node has a crucial influence on the subsequent node. In multiple entity sequences, there may be a connection between the antecedent of one entity and the consequent of another sequence. This characteristic presents challenges for research in association rule discovery on multiple time series data but also offers insights, suggesting that by incorporating some theories and findings from time series analysis, time series data can be transformed to construct new data structures suitable for classic association rule discovery algorithms.

One school of thought in time series research views sequence data from a trend perspective. Williams first proposed the general concept of segments in the field of qualitative reasoning in 1986[9], and in 1990, Cheung and Stephanopoulos formally introduced a trend description language known as “primitivers” [10]. Subsequent research on the feature representation of time series has developed into major schools of thought such as non-adaptive methods, adaptive methods, and model-based methods. Gao and Gu proposed a trend recognition method that converts each node in the sequence data into a polynomial function,

transforming longer time series data into a collection of functions with business significance, providing a new approach and method [11].

Based on the aforementioned research results, this paper introduces the methodology of trend knowledge discovery into the research of association rule discovery in time series data. It fully considers the impact of the lag phenomenon and proposes a method to expand the scope of frequent pattern recognition by calculating the optimal lag coefficient. Considering trend events and lag coefficients can significantly improve the accuracy of predictive models, especially when forecasting market demand, customer behaviour, and economic changes. Through in-depth analysis of time series data, managers can make more precise predictions, thereby optimizing resource allocation and strategic planning. Through in-depth analysis of trend events and lag coefficients, companies can gain a competitive edge in a fiercely competitive market, predict competitors' behaviour, and make corresponding adjustments. Especially in fast-paced industries such as technology and consumer goods, trend analysis and the management of lag effects can help companies maintain their market leadership.

In the field of management, trend events and lag coefficients in time series association rule discovery have significant theoretical and practical implications. Trend events provide managers with tools to identify and predict long-term trends, while lag coefficients help managers understand the delayed effects and causal relationships of decisions. When combined, both can significantly enhance the accuracy and foresight of decision-making, helping companies make more scientific and efficient decisions in complex and uncertain environments.

Therefore, this paper's research has expanded the ideas and methods of association rule discovery on time series data. The method of using trend events to complete data dimensionality reduction is a lossless dimensionality reduction without losing the key features of the sequence data. It is A useful attempt to balance computational cost and algorithm accuracy.

CONCEPTS AND ALGORITHMS

Time Events and The Discretization of Time Series

Discretization is the process of converting continuous-time signals into discrete-time ones, which commonly occurs in fields such as signal processing, statistics, and machine learning. During discretization, factors such as sampling rate, time step size, and how to handle events with different time intervals need to be considered [12]. Many traditional methods have been proposed to address the core issue of discretizing continuous time series data [13]. These methods include the equal-width interval method, the equal-frequency interval method, and the interval merging method, and many others. Some researchers provide a survey of event-driven time series forecasting methods, exploring how to extract useful information from time series with irregular event occurrences and model and predict them. The processing of event data has become an important research topic, especially in fields such as finance, healthcare, and the Internet of things. A generalized framework is proposed to handle event-driven time series data, and a systematic analysis of modelling discretized time series is provided. It offers a methodology to address how to convert irregular time events into discrete time series. Regardless of the method employed, each is based on the original value range of the data and performs certain divisions, thereby possessing inherent limitations. For instance, the equal-width interval method assumes that every possible value of the data contributes equally to the association rules; whereas the equal-frequency interval method struggles to handle new data points that fall outside the current distribution of existing data.

In recent years, some scholars have redefined the concept of “trend” as a polynomial function representation of the stable and predictable components within time series data. By segmenting time series data into multiple trend functions, different nodes within a continuous time series can be distinguished as either trend components or noise components. By discarding the noise nodes and retaining the trend nodes, and then replacing the data nodes with trend functions in the original sequence order to construct a trend sequence, significant dimensionality reduction and noise reduction in high-dimensional time series data can be achieved.

Many traditional methods have been proposed to address the core issue of discretizing continuous time series data. These methods include the equal-width interval method, the equal-frequency interval method, and the interval merging method, and many others. Regardless of the method employed, each is based on the original value range of the data and performs certain divisions, thereby possessing inherent limitations. For instance, the equal-width interval method assumes that every possible value of the data contributes equally to the association rules; whereas the equal-frequency interval method struggles to handle new data points that fall outside the current distribution of existing data.

Trend Events: If a time series $S =< p_1, p_2, \cdots, p_n >$ can be represented by a trend series $< T_1, T_2, \cdots, T_m >$, in which T_i is constructed by nodes between p_a and p_b , then T_i is a trend event on $p_c, p_c \in [p_a, p_b]$.

According to the concept of trend sequences, all data nodes in a specific sequence can be mapped to one or more trend events, i.e., the sequence $S =< p_1, p_2, \cdots, p_n >$ can be transformed into $S^T =< \{T_1\}, \{T_2\}, \cdots, \{T_n\} >$, where $\{T_i\}$ is the set of all trends at a specific node.

However, the trend event sequence constructed through this transformation does not significantly reduce the dimensionality compared to the original sequence. The job is done with trend clustering. At this point, clustering of trend events is used to reduce the numerous trends to a few typical trends. Specifically, hierarchical clustering is used to summarize the trends into several clusters, and then the cluster labels replace all the trend events, thereby as follows Table 1:

Table 1. Algorithm to discretise trend events sequence

IN: Trend Events Sequence TS. OUT: Discretized Trend Event Sequence TS'.
1) create an empty matrix dist with same amount of rows and columns as TS; 2) get T_i from TS; 3) get T_j from TS, $i < j \leq$ length (TS); 4) set dist (i, j) = MWD(T_i, T_j); 5) if j=length (TS), goto 6), else goto 3); 6) if i=length (TS), goto 7), else goto 2); 7) do hierarchical clustering on TS according to dist, and mark with A, B, C...; 8) create a new sequenct TS" with the same length as TS; 9) get T_i from TS; 10) add a symbol from step 7) to TS'(i); 11) save TS', and quit.

Compared to traditional methods, the dimensionality reduction approach involving the introduction of trend functions does not simply segment the original sequence. Instead, it transforms and reconstructs the sequence based on the inherent developmental characteristics embedded within the data. The dimensionality reduction effect is thus highly significant. Furthermore, in the process of constructing trend events, the original sequence is effectively piecewise fitted, laying a solid foundation for subsequent computations, classifications, and predictions.

Lag Coefficients in Time Series Association Rules Discovery

In traditional transactional association rule discovery, the sequence of transactions is generally not considered, or the window used to determine the correlation between transactions is very small. When it is necessary to identify frequent patterns among multiple time series data, there is a possibility that related antecedents and consequents occur asynchronously, a phenomenon known as "lag." Due to the impact of the lag phenomenon, traditional association rule discovery methods, when applied to time series data, might miss important frequent patterns.

In the research of association rules discovery from multiple time series, the transactions are no longer boolean as default but with particular length with start point and end point so that the relationship between antecedents and consequents can be summarized into four cases as shown in the Figure 1. The type marked as (a) represents that a consequent starts after the antecedent totally finished. The traditional study on single time series frequency patten discovery usually work on this type of data. The type marked as (b) shows a data type in which a consequent starts before the antecedent finished. Type (c) is a special type that the antecedent and the consequent start and finish at the same time, in a word, they coincide, which is common in traditional association rules discovery study. The last type marked as (d) shows a much more special kind of data, in which the consequent starts after the antecedent stars but finishes before it finishes, which means the consequent is in the antecedent. In the four types, (b) and (d) are overlooked by the traditional study or they are difficult to solve with classic algorithms.

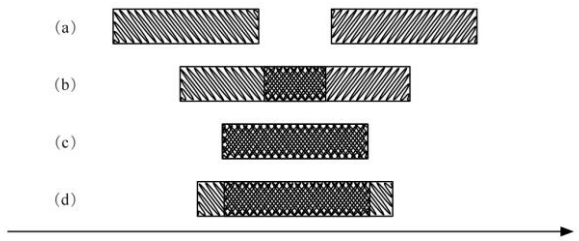

Figure 1. Typical time lag cases

In the research of association rule discovery in time series data, theoretically, a trend antecedent can influence a trend consequent infinitely far in the future. However, this characteristic has no practical significance for solving real-world problems and only increases computational complexity. For time series association rule mining, the influence of the antecedent on the consequent should be confined to a certain interval. Beyond this range, it can be assumed that there is no correlation between the two transactions. Algorithms such as MOWCATL consider this lag phenomenon between data and provide feasible solutions for association rule discovery under this premise. However, they do not discuss how to determine an optimal effective window range for the association. In this paper, this range is defined as the lag coefficient.

Lag Coefficients: For any trend $T = (a, b, M(x))$ in a sequence, when it is used as the consequent of an association rule, any trend that is related to the development and change process of this trend can be represented as $T' = (a', b', M'(x))$, in which $a' \in [a - \tau, a], b' \in (a', b + \tau)$, the τ is the lag coefficient of trend T.

The fact that one trend is within the lag interval of another trend does not necessarily imply a correlation between the two trends. Being within the lag interval is a necessary but not sufficient condition for the existence of a correlation between two trends.

The selection of the lag coefficient is highly flexible. To improve computational efficiency and avoid endless iterations, this paper suggests that the optimal lag coefficient can be determined by combining both subjective and objective methods. Subjectively, the specific business context of the time series data should be considered. For example, if the data is monthly, the upper limit for calculating the lag coefficient can be set to 12, and for quarterly data, the upper limit can be set to 4, and so on.

Based on the subjectively set upper limit, the possible lag coefficients between any two sequences can be further solved by iteration, and the optimal lag coefficient can then be obtained by calculating the average value.

Thus this paper proposed an algorithm whose steps is described as follows Table 2:

Table 2. Algorithm to calculate the best lag coefficients

IN: Trend Events Sequence TS1 and TS2; Max lag coefficient τ_{max}. OUT: Best lag coefficients τ.
1) get a trend $T_1 = (a_1, b_1, M_1(x))$ from TS1; 2) create a new list S_τ, set mwd=0, τ_{tmp}=0; 3) get a trend $T_2 = (a_2, b_2, M_x(x))$ from TS2; 4) if $a_2 < a_1 - \tau_{max}$, and $b_2 > b_1 + \tau_{max}$, set $\tau_{tmp} = \tau_{max}$,and add to S_τ; 5) if $a_1 - \tau_{max} \leq a_2 < a_1$, and $b_2 \leq b_1 + \tau_{max}$, go to 6), else go to 3); 6) if $\lvert MWD(T_1, T_2)\rvert > mwd$, set $mwd = \lvert MWD(T_1, T_2\rvert)$, and go to 7); 7) if $a_1 < a_2 \wedge b_1 < b_2$, then set $\tau_{tmp} = a_1 - a_2$; 8) if $a_1 > a_2 \wedge b_1 < b_2$, then set $\tau_{tmp} = \max(a_1 - a_2, b_2 - b_1)$; 9) add τ_{tmp} to S_τ, 10) if more trend left in TS2, go to 3), else go to 11); 11) if more trend left in TS1, go to 1), else go to 12); 12) set $\tau = \lceil \overline{S_\tau} \rceil$ 13) save τ, and quit.

After obtaining the optimal lag coefficient, it can be used to compare all transactions, marking them as T (True) or F (False), thereby achieving the binarization of time series data.

EXPERIMENTS

Dataset and Pre-Processing Steps

(1) In this paper, we select the publicly available dataset "PM2.5 Data of Five Chinese Cities" from UCI, which includes PM2.5 and related meteorological data for five cities: Beijing (BJ), Shanghai (SH), Guangzhou (GZ), Chengdu (CD), and Shenyang (SY).

(2) Due to the presence of a large number of missing values in the original dataset and the enormous scale of the data, some preprocessing is necessary. The key steps of preprocessing in this paper include:

(3) Only considering data from January 1, 2014, onwards to avoid the issue of numerous missing values in the earlier data.

(4) Using only the data released by the U.S. Embassy in the five cities as samples to ensure consistent data accuracy.

(5) Retaining only the daily data recorded at 12 AM to reduce the data scale and enhance computational efficiency.

(6) Removing all dimensions where the data is entirely zero to reduce interference. For the dimensions "precipitation" and "iprec," which contain a large number of zero values, manually transforming them into two values, "0" and "non-0," to achieve discretization.

Key Steps and the Result of the Algorithms

After completing the preprocessing, the first step of the experiment is to conduct trend identification to construct trend sequences. Using the trend identification algorithm proposed by Gao and Gu, most dimension sequences are set with an error threshold of 0.2. For HUMI, due to the particular nature of the data, the threshold is set at 0.05. The basic situation of the dataset used to test the algorithm in this paper is shown in the Table 3 after processing.

Table 3. PM2.5 Data of Five Chinese Cities Data after pre-Process

City	Amount of Affairs	Amount of Trend					
		PM2.5	DEWP	TEMP	HUMI	PRES	IWS
BJ	685	93	179	151	148	256	56
CD	647	132	250	192	136	287	75
GZ	695	132	428	293	107	433	86
SH	1004	152	309	282	111	503	106
SY	600	76	127	146	81	208	55

We focus on PM2.5 data, thus setting it as the target dimension for constraint, specifically as the sole consequent in association rules. All dimensions undergo trend clustering, and using the cluster labels as trend events, each dimension is transformed into a boolean dataset of trend events.

Continuing with the processed dataset mentioned above, considering it is daily data, a maximum lag coefficient of 7 is chosen. For association rule discovery, a support threshold of 0.1 and confidence threshold of 0.5 are applied. The classic Apriori algorithm is then used to obtain association rules. Due to the large number of rules generated, to save space, only the top 5 rules from each city's results are listed in the Table 4.

Table 4. Top Rules Discoverd in Each City

City	lhs	rhs	supp	coefi	lift	count
BJ	{pres-4, temp-7}	pm-6	0.1051	0.9118	2.2414	31
	{humi-10, pres-4, temp-7}	pm-6	0.1051	0.9118	2.2414	31
	{pres-1, pres-4, temp-7}	pm-6	0.1051	0.9118	2.2414	31
	{humi-10, pres-1, pres-4, temp-7}	pm-6	0.1051	0.9118	2.2414	31
	{pres-4, SE, temp-7}	pm-6	0.1017	0.9091	2.2348	30
CD	{humi-13, humi-16, pres-5, SW}	pm-13	0.1057	0.7167	4.0512	43
	{humi-13, humi-15, humi-16, pres-5, SW}	pm-13	0.1057	0.7167	4.0512	43
	{humi-13, humi-16, humi-6, pres-5, SW}	pm-13	0.1057	0.7167	4.0512	43
	{humi-13, humi-16, humi-2, pres-5, SW}	pm-13	0.1057	0.7167	4.0512	43
	{humi-13, humi-16, humi-8, pres-5, SW}	pm-13	0.1057	0.7167	4.0512	43
GZ	{dewp-10, dewp-4, dewp-9, pres-4}	pm-12	0.1042	0.7018	3.2863	40
	{dewp-10, dewp-4, dewp-5, dewp-9, pres-4}	pm-12	0.1042	0.7018	3.2863	40
	{dewp-10, dewp-2, dewp-4, dewp-9, pres-4}	pm-12	0.1042	0.7018	3.2863	40
	{dewp-10, dewp-4, dewp-9, humi-16, pres-4}	pm-12	0.1042	0.7018	3.2863	40
	{dewp-10, dewp-4, dewp-9, humi-18, pres-4}	pm-12	0.1042	0.7018	3.2863	40
SH	{dewp-1, humi-12, lws-1, pres-3}	pm-5	0.1011	0.8689	1.8813	53
	{dewp-1, humi-12, lws-1, pres-2, pres-3}	pm-5	0.1011	0.8689	1.8813	53
	{dewp-1, humi-12, lws-1, NE, pres-3}	pm-5	0.1011	0.8689	1.8813	53
	{dewp-1, humi-11, humi-12, lws-1, pres-3}	pm-5	0.1011	0.8689	1.8813	53
	{dewp-1, humi-12, lws-1, pres-1, pres-3}	pm-5	0.1011	0.8689	1.8813	53
SY	{NW, SE, SW}	pm-2	0.1137	0.5	1.4655	29
	{NW, pres-1, SE, SW}	pm-2	0.1059	0.5	1.4655	27
	{dewp-5}	pm-1	0.1098	1.0	1.4167	28
	{dewp-2}	pm-1	0.1255	1.0	1.4167	32
	{dewp-3}	pm-1	0.1333	1.0	1.4167	34

EXPERIMENTS RESULT ANALYSIS

From the above results, it is evident that each of the five cities exhibits distinctive factors related to PM2.5, and the implicit correlations also reflect typical regional characteristics.

Among the five cities, Beijing has a temperate monsoon climate and is significantly affected by air pressure, temperature and humidity. In contrast, Shanghai is located in the subtropical estuary of the Yangtze River, which has more prominent oceanic characteristics. At this time, if the air pressure is low and the wind speed is fast, PM2.5 will be less likely to accumulate. Chengdu is located in an inland basin, and the humid climate makes the wind direction and wind speed relatively stable. The factors that have the greatest impact on PM2.5 are humidity and air pressure, which determine precipitation. Guangzhou is more like a combination of Shanghai and Chengdu. As a tropical coastal city, the dew point indicator that reflects the balance of moisture and temperature in the air is the most important for Guangzhou. The most special of all cities is Shenyang, which has the highest latitude and the lowest average temperature. It should be significantly affected by factors such as air pressure, but this is not the case. Due to the humid water vapor brought by the monsoon in the Bohai Bay, Shenyang's PM2.5 performance becomes similar to that of Guangzhou, and the dew point plays an important role.

The analysis above shows that in practical applications, it is crucial to analyze each specific situation to identify the primary influencing factors effectively, thereby achieving the goal of controlling and adjusting PM2.5 levels.

At the same time, the results of the experiment demonstrate that through the methods presented in this paper, hidden association rules in time series data can be successfully identified, indicating both theoretical and practical significance.

CONCLUSION

This paper re-examines the problem of association rule discovery on time series data from the perspective of trend knowledge discovery. It introduces the concept of trend events to discretize continuous sequence data and fully considers the unique lag issues faced when conducting association rule discovery on time series data. By solving for the optimal lag coefficient, it further discretizes the sequence data, enabling traditional association rule discovery algorithms to operate effectively. Experimental validation on publicly available datasets demonstrates the feasibility of the proposed method and its capability to represent real-world problems. This work represents a beneficial attempt in the research of association rule discovery algorithms.

However, the research method given in this article has high computational complexity and high computational cost, and the results obtained also have a certain degree of ambiguity. When making specific interpretations and explanations, the business background of the research needs to be fully considered in order to be more convincing conclusion. On the other hand, the trend discovery of sequence data is the basis of this article, but the effect of trend identification is closely related to the sampling frequency of sequence data. How to balance the computational complexity and algorithm accuracy to achieve dynamic and automatic analysis granularity determination will It is one of the important research directions in the future.

REFRENCES

[1] Agrawal, R., Faloutsos, C., Swami, A., Efficient similarity search in sequence databases, Springer, (1993).
[2] Han, J., Pei, J., Mortazavi-Asl, B., Chen, Q., Dayal, U., Hsu, M., FreeSpan: frequent pattern-projected sequential pattern mining, (2000).
[3] Han, J., Pei, J., Mortazavi-Asl, B., Pinto, H., Chen, Q., Dayal, U., et al., Prefixspan: Mining sequential patterns efficiently by prefix-projected pattern growth, (2001).
[4] Agrawal, R., Srikant, R., Mining sequential patterns, Proceedings of the Eleventh International Conference on Data Engineering, 1995.
[5] Zaki, M. J., SPADE: An efficient algorithm for mining frequent sequences, Machine learning (2001), 42(1), 31-60.
[6] S. Weitl-Harms, J. Deogun and T. Tadesse, Discovering sequential association rules with constraints and time lags in multiple sequences.
[7] Yang Ren, Chidambaram, M, Closed-loop identification of second-order plus time delay (SOPTD) model of multivariable systems by optimization method, Industrial & Engineering Chemistry Research (2012), 51(28), 9620-9633.
[8] Jiannan, Cai., Qiliang, L., Feng, X., Min, D., Zhanjun, H., Jianbo, T., Multi-level spatial homolocation pattern adaptive mining method, Acta Geodaetica et Cartographica Sinica (2016), 4, 11.
[9] Williams, B. C., Doing Time: Putting Qualitative Reasoning on Firmer Ground, (1986).
[10] Cheung, J., Stephanopoulos, G., Representation of process trends—Part I. A formal representation framework, Computers & Chemical Engineering (1990), 14(4), 495-510.
[11] Xuedong, G., Ka, G., An algorithm for trend partition detection from sequential data, International Journal of Computer Science & Applications (2016), 13(1).
[12] Robert H. Shumway and David S. Stoffer.Time Series Analysis and Its Applications: With R Examples, Springer ,2017.
[13] Peter J. Brockwell and Richard A. Davis. Introduction to Time Series and Forecasting.Springer,2016.