An approach to discovering multi-temporal patterns and its application to financial databases

ARTICLE *in* INFORMATION SCIENCES · MARCH 2010

Impact Factor: 4.04 · DOI: 10.1016/j.ins.2009.08.026 · Source: DBLP

3 AUTHORS, INCLUDING:

Qiang Wei
Tsinghua University
46 PUBLICATIONS **474** CITATIONS

SEE PROFILE

Guoqing Chen
State Key Laboratory Of Geohazard Pr…
126 PUBLICATIONS **1,223** CITATIONS

SEE PROFILE

Information Sciences 180 (2010) 873–885

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

An approach to discovering multi-temporal patterns and its application to financial databases

Xiaoxiao Kong, Qiang Wei *, Guoqing Chen

Department of Management Science and Engineering, School of Economics and Management, Tsinghua University, Beijing 100084, China

ARTICLE INFO

Article history:
Received 18 March 2009
Received in revised form 7 August 2009
Accepted 29 August 2009

Keywords:
Data mining
Multi-temporal pattern
Association rule
Associative financial movement

ABSTRACT

Managerial decision-making processes often involve data of the time nature and need to understand complex temporal associations among events. Extending classical association rule mining approaches in consideration of time in order to obtain temporal information/knowledge is deemed important for decision support, which is nowadays one of the key issues in business intelligence. This paper presents the notion of multi-temporal patterns with four different temporal predicates, namely *before, during, equal* and *overlap*, and discusses a number of related properties, based on which a mining algorithm is designed. This enables us to effectively discover multi-temporal patterns in large-scale temporal databases by reducing the database scan in the generation of candidate patterns. The proposed approach is then applied to stock markets, aimed at exploring possible associative movements between the stock markets of Chinese mainland and Hong Kong so as to provide helpful knowledge for investment decisions.

1. Introduction

In recent years, discovery of association rules [1,3,14,19,21,31] and sequential patterns [2,5,6,11,12,15,20,23–26,32,34] has been a major research issue in the area of data mining and knowledge discovery. Typical association rules usually reflect related events occurring at the same time, and sequential patterns represent commonly occurring sequences that are in a time order. However, real-world businesses often generate a massive volume of data in daily operations and decision-making processes, which are of a richer temporal nature. Especially in financial markets (e.g., stock markets), the temporal nature of data is a key factor in understanding the dynamics of markets. For instance, stock A's price increases *during* the period when stock B's price decreases. Stock C's price increases *before* the Market index increases. The exchange rates of USD/CNY and USD/HKD change in the same/*equal* period. Usually, these patterns may not appear in 100% of the cases, but sufficiently in a frequent and significant manner. Apparently, such temporal relationships (e.g., *during, before* and *equal*) are certain kinds of the real-world semantics that are considered meaningful and useful in practice. If the hidden temporal patterns in massive financial databases could be effectively discovered, then the dynamics of the financial markets could be well described. In fact, finance is one of the fields where data mining methods have been widely adopted to support decision-making.

Generally, temporal relationships between events with different time stamps could be categorized into several types in forms of temporal comparison predicates such as *after, before, meet, overlap, during, start, finish* and *equal* [4]. Though recent years have witnessed certain efforts in discovering the *after*/*before* relationship [2,8–10,24,26,36], the investigation of other temporal relationships is still badly needed. A few explorations in this regard then include finding temporal relationships

* Corresponding author. Tel.: +86 10 62789824; fax: +86 10 62785876.
E-mail address: weiq@sem.tsinghua.edu.cn (Q. Wei).

0020-0255/$ - see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.ins.2009.08.026

with time-interval-based events by Rainsford and Roddic [29], Hoppner [13], Giannotti et al. [12], Winarko and Roddick [32] and Wu and Chen [34], with mutually delayed events by Yu and Chen [35], with during-events by Zhang et al. [37], and other extensions [15,20]. These discovered temporal patterns are, however, represented in single predicates, such as *during*. The effort in discovering temporal patterns with multiple predicates (e.g., both *during* and *before*) is still quite limited [5,7,18] but worthy (e.g., stock A's price increases *during* the period when stock B's price increases, where stock B's price increases *before* stock C's price increases), and open for in-depth exploration.

This paper is organized as follows: Section 2 will introduce some preliminaries of temporal databases, temporal predicates, multi-temporal patterns as well as the concepts of degrees of support and confidence. Some properties of multi-temporal patterns will be discussed in Section 3. Based on the notions and properties of multi-temporal patterns, the mining algorithm will be proposed in Section 4. In Section 5, the proposed algorithm will be applied to Chinese mainland and Hong Kong stock markets so as to test possible associative movements between these two markets.

2. Multi-temporal patterns

Objects in the real world (e.g., people, machines, plants, etc.) can be represented by their attributes. A value of an attribute for an object reflects a feature state of the object. In other words, we may use a triplet $s = \langle o, a, v\rangle$ to represent a state, where o is an object, a is an attribute of o, and v is a value of attribute a. For example, such a state may look like ⟨stock A, price, increase ⟩.

In many cases, a state takes place in a certain time interval, which is called an event denoted as e, and can be represented using another triplet $e = \langle s, st, et\rangle$, where s is a state, st is the start time of the state and et is the end time of the state. For instance, with s=⟨stock A, price, increase⟩, e=$\langle s, 1, 3\rangle$ means that "stock A's price increases in time interval [1,3]".

Suppose S is a set of n states, i.e., $S = \{s_1, s_2, \ldots, s_n\}$. Each state $s_i (s_i \in S, 1 \leqslant i \leqslant n)$ can correspond to several events, e_{ij}, where $j = 1, 2, \ldots, m_i$. Then a temporal database, denoted as D_T, can be represented as a set of events sorted by start time in an ascending order. Table 1 illustrates an example of a temporal database D_T.

Definition 1. For an event e_A and its time interval (st_A, et_A), given a pre-set time lag called *window*, then C_A=[st_A, et_A+*window*] is defined as the *window* constraint of e_A. Given another event e_B, if $st_B \in C_A$, then e_B is called to satisfy the *window* constraint of e_A.

Window constraint is used as an interestingness measure, since if two events are far apart from each other to some extent, then these two events will not be regarded associated semantically. This constraint is quite useful in many situations [9].

Definition 2. Given two events, e_A and e_B, with corresponding time intervals (st_A, et_A) and (st_B, et_B), respectively, then these two events are called to compose a temporal instance if they are connected by one of the four temporal predicates (i.e., *equal, before, during* or *overlap*) as follows:

(1) If $st_A = st_B$ and $et_A = et_B$, then e_A *equal* e_B;
(2) If $0 \leqslant st_B - et_A \leqslant$ *window*, then e_A *before* e_B;
(3) If $st_B < st_A < et_A \leqslant et_B$ or $st_B \leqslant st_A < et_A < et_B$, then e_A *during* e_B;
(4) If $st_A \leqslant st_B < et_A < et_B$ or $st_A < st_B < et_A \leqslant et_B$, then e_A *overlap* e_B.

Notably, *equal, before, during* and *overlap* are considered to be the most popular temporal predicates [4]. For simplicity, we will denote *equal, before, during* and *overlap* as E, B, D and O. Then a temporal instance between two events e_A and e_B can be denoted as $\phi := e_A \overset{R}{\Rightarrow} e_B$, where R is a temporal predicate, i.e., $R \in \{D, B, O, E\}$. For example, in Table 1, we have $e_{31} \overset{D}{\Rightarrow} e_{11}, e_{61} \overset{B}{\Rightarrow} e_{22}$ and $e_{62} \overset{O}{\Rightarrow} e_{42}$.

Furthermore, due to the fact that a temporal instance reflects a certain temporal relationship between two states (e.g., s_i and s_j), finding whether such a relationship holds in other temporal instances is of great interest, as this relationship repre-

Table 1
An example of a temporal database D_T.

Event	State	st	et	Event	State	st	et
e_{11}	s_1	1	5	e_{32}	s_3	33	38
e_{31}	s_3	1	4	e_{62}	s_6	34	39
e_{51}	s_5	2	10	e_{42}	s_4	25	37
e_{21}	s_2	2	8	e_{52}	s_5	27	42
e_{61}	s_6	2	5	e_{23}	s_2	28	32
e_{22}	s_2	4	6	e_{12}	s_1	30	40
e_{41}	s_4	3	7	e_{33}	s_3	30	38
...	...	...	...	...	...	...	...

sents a temporal pattern $s_i \overset{R}{\Rightarrow} s_j$. Concretely, for a temporal pattern $s_i \overset{R}{\Rightarrow} s_j$, if there exists a temporal instance, e.g., $e_{ip} \overset{R}{\Rightarrow} e_{jq}$, where e_{ip} and e_{jq} are the events with respect to s_i and s_j, respectively, then we call the temporal pattern $s_i \overset{R}{\Rightarrow} s_j$ to be supported by the temporal instance $e_{ip} \overset{R}{\Rightarrow} e_{jq}$. For example, in Table 1, the temporal pattern $s_1 \overset{O}{\Rightarrow} s_2$ is supported by $e_{11} \overset{O}{\Rightarrow} e_{21}, e_{11} \overset{O}{\Rightarrow} e_{22}$, and $e_{12} \overset{O}{\Rightarrow} e_{23}$. Thus, from the viewpoint of data mining and knowledge discovery, only the temporal patterns frequently supported by temporal instances will be discovered as useful patterns for decision-making.

Definition 3. Given a set of n states S, e.g., $S = \{s_1, s_2, \ldots, s_n\}$, and the set of temporal predicates $\{D, B, O, E\}$, then a pattern of $k+1$ states connected with k temporal predicates is called a temporal pattern with length = k (the number of predicates is k), or $(k+1)$-state (the number of states is $k+1$) temporal pattern, denoted as Φ_k. Then we have:

(1) When $k = 0, \Phi_0 := s_0, s_0 \in S$ (we call s_0 a degenerated temporal pattern);
(2) When $k = 1, \Phi_1 := (\Phi_0 \overset{R_1}{\Rightarrow} s_1) := (s_0 \overset{R_1}{\Rightarrow} s_1), R_1 \in \{D, B, O, E\}$;
(3) When $k = 2, \Phi_2 := (\Phi_1 \overset{R_2}{\Rightarrow} s_2) := (s_0 \overset{R_1}{\Rightarrow} s_1 \overset{R_2}{\Rightarrow} s_2), R_1, R_2 \in \{D, B, O, E\}$;
(4) When $k > 2, \Phi_k := (\Phi_{k-1} \overset{R_k}{\Rightarrow} s_k) := (s_0 \overset{R_1}{\Rightarrow} s_1 \overset{R_2}{\Rightarrow} s_2 \overset{R_3}{\Rightarrow} \ldots \overset{R_k}{\Rightarrow} s_k), R_j \in \{D, B, O, E\}$, where $j = 1, 2, \ldots, k$.

Clearly, when k=0, the pattern contains only one state, e.g., "stock A's price increase". When k=1, it means that two states have a temporal relationship, e.g., "stock A's price increase" $\overset{B}{\Rightarrow}$ "Shanghai stock market index increase" meaning that stock A's price increases *before* the market index increases. Moreover, when $k \geqslant 2$, it represents several states having sequentially temporal relationships, e.g., "stock A's price increase" $\overset{B}{\Rightarrow}$ "Shanghai stock market index increase" $\overset{B}{\Rightarrow}$ "Shenzhen stock market index increase". Apparently, this temporal pattern reflects certain knowledge that is useful to decision-makers in understanding the dynamics of China stock markets. Note that if the same predicate (such as B) is used in the pattern, the pattern is called a single-temporal pattern. If different predicates are integrated into a temporal pattern, then the pattern is called a multi-temporal pattern. As an example, "stock B's price decrease" $\overset{D}{\Rightarrow}$ "stock A's price increase" $\overset{B}{\Rightarrow}$ "Shanghai market index increase" represents a multi-temporal pattern with two predicates D and B.

Similarly to the patterns and instances with two states, a multi-temporal pattern $s_0 \overset{R_1}{\Rightarrow} s_1 \overset{R_2}{\Rightarrow} s_2 \overset{R_3}{\Rightarrow} \ldots \overset{R_k}{\Rightarrow} s_k$ could be supported by multiple temporal instances, e.g., $e_{0p0} \overset{R_1}{\Rightarrow} e_{1p1} \overset{R_2}{\Rightarrow} e_{2p2} \overset{R_3}{\Rightarrow} \ldots \overset{R_k}{\Rightarrow} e_{kpk}$. For example, in Table 1, the temporal pattern $\Phi := s_3 \overset{D}{\Rightarrow} s_1 \overset{O}{\Rightarrow} s_2$ is supported by the temporal instances: $e_{31} \overset{D}{\Rightarrow} e_{11} \overset{O}{\Rightarrow} e_{21}$ and $e_{33} \overset{D}{\Rightarrow} e_{12} \overset{O}{\Rightarrow} e_{23}$. Intuitively, if a temporal pattern is frequently supported by temporal instances to a certain extent, then the pattern will be regarded as a qualified pattern. In order to discover qualified patterns, the notion of effective time interval is introduced as follows.

Definition 4. Given a temporal pattern $\Phi, \Phi := s_0 \overset{R_1}{\Rightarrow} s_1 \overset{R_2}{\Rightarrow} s_2 \overset{R_3}{\Rightarrow} \ldots \overset{R_k}{\Rightarrow} s_k, R_j \in \{D, B, O, E\}, j = 1, 2, \ldots, k$, with a supporting temporal instance $\phi, \phi := e_{0p0} \overset{R_1}{\Rightarrow} e_{1p1} \overset{R_2}{\Rightarrow} e_{2p2} \overset{R_3}{\Rightarrow} \ldots \overset{R_k}{\Rightarrow} e_{kpk}$, then the effective time interval for ϕ supporting Φ, denoted as $t_{\Phi\phi}$, is defined as:

$$t_{\Phi\phi} = t_{kpk} \cup t_{(k-1)p(k-1)} \cup \ldots \cup t_{1p1} \cup t_{0p0},$$

where $t_{ipj} = (st_{ipj}, et_{ipj})$ is the time interval of event $e_{jpj}, j = 0, 2, \ldots, k$. Then $t_{\Phi\phi} = (st_{\Phi\phi}, et_{\Phi\phi})$, where $st_{\Phi\phi} = \min\{st_{jpj} | j = 0, 2, \ldots, k\}, et_{\Phi\phi} = \max\{et_{jpj} |\ j = 0, 2, \ldots, k\}$. Furthermore, $\mathrm{T}(\Phi) = \{t_{\Phi\phi} | t_{\Phi\phi}$ is the effective time interval for temporal instance ϕ supporting $\Phi\}$ is called the set of effective time intervals for all the temporal instances supporting temporal pattern Φ.

For example, in Table 1, the temporal pattern $\Phi := s_3 \overset{D}{\Rightarrow} s_1 \overset{O}{\Rightarrow} s_2$ is supported by the temporal instance $\phi := e_{31} \overset{D}{\Rightarrow} e_{11} \overset{O}{\Rightarrow} e_{21}$, then the effective time interval for ϕ supporting Φ is $t_{\Phi\phi} = (st_{\Phi\phi}, et_{\Phi\phi}) = (min\{1, 1, 2\}, max\{4, 5, 8\}) = (1, 8)$.

Based on the notion of effective time interval, the following properties could be obtained.

Theorem 1. *Given a temporal pattern $\Phi, \Phi := s_0 \overset{R_1}{\Rightarrow} s_1 \overset{R_2}{\Rightarrow} s_2 \overset{R_3}{\Rightarrow} \ldots \overset{R_k}{\Rightarrow} s_k, R_j \in \{D, B, O, E\}, j = 1, 2, \ldots, k$, with a supporting temporal instance $\phi, \phi := e_{0p0} \overset{R_1}{\Rightarrow} e_{1p1} \overset{R_2}{\Rightarrow} e_{2p2} \overset{R_3}{\Rightarrow} \ldots \overset{R_k}{\Rightarrow} e_{kpk}$, the effective time interval for ϕ supporting Φ, i.e., $t_{\Phi\phi}$ has the following properties:*

(1) $st_{jpj} \geqslant st_{\Phi\phi}, et_{ipj} \leqslant et_{\Phi\phi}$, and $et_{kpk} = et_{\Phi\phi}$, where $j = 0, 2, \ldots, k$;
(2) If $R_1 = R_2 = \ldots = R_k = E$, then $st_{\Phi\phi} = st_{1p1} = st_{1p2} = \ldots = st_{kpk}$, and $et_{\Phi\phi} = et_{1p1} = et_{1p2} = \ldots = et_{kpk}$;
(3) If $R_1 = R_2 = \ldots = R_k = D$, then $st_{\Phi\phi} = st_{kpk}$, and $et_{\Phi\phi} = et_{kpk}$;
(4) If $R_1 = R_2 = \ldots = R_k = B$ or D, then $st_{\Phi\phi} = st_{1p1}$, and $et_{\Phi\phi} = et_{kpk}$;

where the time interval of event e_{jpj} corresponding to state s_j is t_{jpj}=(st_{jpj}, et_{jpj}).

Proof

(1) According to Definition 4, we have $st_{\Phi\phi} = \min\{st_{0p0}, st_{1p1}, \ldots, st_{kpk}\}, et_{\Phi\phi} = \max\{et_{0p0}, et_{1p1}, \ldots, et_{kpk}\}$. So $st_{jpj} \geqslant st_{\Phi\phi}$, $et_{ipj} \leqslant et_{\Phi\phi}, j = 0, 1, 2, \ldots, k$.

For the temporal pattern $\Phi_1 := s_0 \overset{R_1}{\Rightarrow} s_1$ supported by $\phi_1 := e_{0p0} \overset{R_1}{\Rightarrow} e_{1p1}$ according to Definition 2, we have $et_{0p0} \leqslant et_{1p1}$, then according to Definition 4, $et_{\Phi 1\phi 1} = et_{1p1}$. Similarly, for the temporal pattern $\Phi_2 := \Phi_1 \overset{R_2}{\Rightarrow} s_2$ supported by $\phi_1 := e_{0p0} \overset{R_1}{\Rightarrow} e_{1p1} \overset{R_2}{\Rightarrow} e_{2p2}$, we have $et_{0p0} \leqslant et_{1p1} \leqslant et_{2p2}$, then $et_{\Phi 2\phi 2} = et_{2p2}$. And so on and so forth, for the temporal pattern $\Phi := s_0 \overset{R_1}{\Rightarrow} s_1 \overset{R_2}{\Rightarrow} s_2 \overset{R_3}{\Rightarrow} \ldots \overset{R_k}{\Rightarrow} s_k = \Phi_{k-1} \overset{R_k}{\Rightarrow} s_k$, we have $et_{0p0} \leqslant et_{1p1} \leqslant \ldots \leqslant et_{kpk}$, then $et_{\Phi\phi} = et_{kpk}$.

(2) It can be proved directly from Definition 2 (1).
(3) Based on Definition 2 (3), it can be inferred that, $et_{0p0} \leqslant et_{1p1} \leqslant \ldots \leqslant et_{kpk}$, and $st_{kpk} \leqslant st_{(k-1)p(k-1)} \leqslant \ldots \leqslant st_{0p0}$. Since $st_{\Phi\phi} = \min\{st_{0p0}, st_{1p1}, \ldots, st_{kpk}\}, et_{\Phi\phi} = \max\{et_{0p0}, et_{1p1}, \ldots, et_{kpk}\}$, then $st_{\Phi\phi} = st_{kpk}$, and $et_{\Phi\phi} = et_{kpk}$.
(4) According to Definition 2 ((2) and (4)), it can be inferred that, $st_{0p0} < st_{1p1} < \ldots < st_{kpk}$. Since $st_{\Phi\phi} = \min\{st_{0p0}, st_{1p1}, \ldots, st_{kpk}\}$, then $st_{\Phi\phi} = st_{1p1}$, and $et_{\Phi\phi} = et_{kpk}$ (According to proof (1)). □

A temporal instance is an instance for its corresponding temporal pattern, reflecting all the states happening in a certain time interval. In order to measure the frequency that the states supported by multiple instances, the degree of support and the degree of confidence are introduced below in a similar spirit to [1–3,14,19,21,31,35,36].

Definition 5. In a temporal database, for a given temporal pattern Φ, the number of temporal instances that support the temporal pattern is denoted as $supp(\Phi)$. Then the degree of support of a temporal pattern Φ is defined as:

$$support(\Phi) = supp(\Phi)/|E|,$$

where $|E| = \max_{j=1,2,\ldots,N}(|E_j|), N$ is the number of states in a given temporal database D_T, E_j is the set of events supporting state $s_j, |E_j|$ is the cardinality of E_j (i.e., the number of the events in E_j), and e_{jp} is an event supporting state $s_j, i.e., e_{jp} \in E_j$.

For example, in Table 1, the multi-temporal pattern $\Phi := s_3 \overset{D}{\Rightarrow} s_1 \overset{O}{\Rightarrow} s_2$ is supported by the temporal instances $e_{31} \overset{D}{\Rightarrow} e_{11} \overset{O}{\Rightarrow} e_{21}$ and $e_{33} \overset{D}{\Rightarrow} e_{12} \overset{O}{\Rightarrow} e_{23}$, which reflects that $\mathrm{supp}(\Phi) = 2$. In Table 1, it can be seen that $E_1 = \{e_{11}, e_{12}\}, E_2 = \{e_{21}, e_{22}, e_{23}\}, E_3 = \{e_{31}, e_{32}, e_{33}\}, E_4 = \{e_{41}, e_{42}\}, E_5 = \{e_{51}, e_{52}\}$ and $E_6 = \{e_{61}, e_{62}\}$. So $|E| = \max(2,3,3,2,2,2) = 3$. Then $support(s_3 \overset{D}{\Rightarrow} s_1 \overset{O}{\Rightarrow} s_2) = 2/3$. The degree of support is used to measure the strength (frequency) of a multi-temporal pattern in a certain temporal database.

Subsequently, the following property could be obtained.

Property 1. *Given any temporal pattern* $\Phi := s_0 \overset{R_1}{\Rightarrow} s_1 \overset{R_2}{\Rightarrow} s_2 \overset{R_3}{\Rightarrow} \ldots \overset{R_k}{\Rightarrow} s_k$, *where* E_j *supports* s_j, *then* $supp(\Phi) \leqslant |E_j| \leqslant |E|$.

Proof. According to Definitions 1 and 3, if an event e supports s_j (i.e., $e \in E_j$), it may be in an event instance to support Φ. But if the event e does not support s_j, then it is not in any event instances to support Φ. So $supp(\Phi) \leqslant |E_j|$.

Moreover, according to Definition 5, we have $|E_j| \leqslant |E|$.

Thus, $supp(\Phi) \leqslant |E_j| \leqslant |E|$. □

Based on Definition 3 and the concepts of association rule [1–3,14,19,21,31,36], considering the semantics of temporal predicates, the degree of confidence for a temporal pattern Φ can be defined as follows in Definition 6.

Definition 6. In a temporal database, for a given temporal pattern $\Phi := s_0 \overset{R_1}{\Rightarrow} s_1 \overset{R_2}{\Rightarrow} s_2 \overset{R_3}{\Rightarrow} \ldots \overset{R_k}{\Rightarrow} s_k$, let Φ_{k-1} be the $(k-1)$ -item sub-pattern of Φ, i.e., $s_0 \overset{R_1}{\Rightarrow} s_1 \overset{R_2}{\Rightarrow} s_2 \overset{R_3}{\Rightarrow} \ldots \overset{R_{k-1}}{\Rightarrow} s_{k-1}$, then Φ can be denoted as $\Phi_{k-1} \overset{R_k}{\Rightarrow} s_k$. Furthermore, the degree of confidence for a temporal pattern Φ is defined as:

$$confidence(\Phi) = \begin{cases} \frac{\sup p(\Phi)}{\sup p(\Phi_{k-1})} \times confidence(\Phi_{k-1}) & R_k \neq D \\ \frac{\sup p(\Phi)}{\sup p(s_k)} \times confidence(s_k) = \frac{\sup p(\Phi)}{\sup p(s_k)} & R_k = D \\ 1 & \Phi \text{ is a degenerated temporal pattern} \end{cases}$$

For example, in Table 1, for any $\Phi := s_i, i = 1, 2, \ldots, N$, we have $confidence(\Phi) = 1$, since Φ is a degenerated temporal pattern and there is no antecedent. For $\Phi := s_3 \overset{D}{\Rightarrow} s_1 (R_k = D, k = 1)$, we have $supp(s_3 \overset{D}{\Rightarrow} s_1) = 2, supp(s_1) = 2$, and $confidence(s_1) = 1$. Then $confidence(s_3 \overset{D}{\Rightarrow} s_1) = supp(s_3 \overset{D}{\Rightarrow} s_1)/supp(s_1) = 2/2 = 100\%$, which means s_3 will occur at a 100% chance during s_1's occurrence. Moreover, for the multi-temporal pattern $\Phi := s_3 \overset{D}{\Rightarrow} s_1 \overset{O}{\Rightarrow} s_2$ $(R_K \neq D, k = 2)$, we have $supp(s_3 \overset{D}{\Rightarrow} s_1 \overset{O}{\Rightarrow} s_2) = 2, supp(s_3 \overset{D}{\Rightarrow} s_1) = 2$, and $confidence(s_3 \overset{D}{\Rightarrow} s_1) = 1$. Then $confidence(s_3 \overset{D}{\Rightarrow} s_1 \overset{O}{\Rightarrow} s_2) = supp(s_3 \overset{D}{\Rightarrow} s_1 \overset{O}{\Rightarrow} s_2)/supp(s_3 \overset{D}{\Rightarrow} s_1) \times confidence(s_3 \overset{D}{\Rightarrow} s_1) = 2/2 \times 100\% = 100\%$. Basically, the degree of confidence is used to evaluate the significance of a multi-temporal pattern in a certain temporal database.

Based on the definitions of the degree of support and the degree of confidence for a multi-temporal pattern, given the thresholds, e.g., the minimal support α and the minimal confidence β, where $\alpha, \beta \in [0,1]$, a pattern Φ with $support(\Phi) \geqslant \alpha$ is called a frequent (candidate) pattern. A pattern Φ with $support(\Phi) \geqslant \alpha$ and $confidence(\Phi) \geqslant \beta$ is called a qualified pattern.

3. Properties of multi-temporal patterns

Before constructing the mining algorithm to discover qualified multi-temporal patterns, some useful properties could be derived and used in further algorithmic design so as to improve the effectiveness.

Definition 7. Given a multi-temporal pattern $\Phi, \Phi := s_0 \overset{R_1}{\Rightarrow} s_1 \overset{R_2}{\Rightarrow} s_2 \overset{R_3}{\Rightarrow} \ldots \overset{R_k}{\Rightarrow} s_k, R_j \in \{D, B, O, E\}, j = 1, 2, \ldots, k$, then a pattern $\Phi', \Phi' := s_{j1} \overset{R_{j2}}{\Rightarrow} s_{j2} \overset{R_{j3}}{\Rightarrow} s_{j3} \ldots \overset{R_{jm}}{\Rightarrow} s_{jm}$, where $0 \leqslant j_1 \leqslant j_2 \leqslant \ldots \leqslant j_m \leqslant k$ and $0 \leqslant m \leqslant k$, is called a m-item sub-pattern of Φ. If $j_1, j_2, \ldots, j_m$ is a series of sequential natural numbers (i.e., $j_p = j_{p-1} + 1, p = 1, 2, \ldots, m$ and $j_1 = 0$), then Φ' is called a sequential sub-pattern of Φ.

Definition 8. Given a temporal pattern Φ and a sub-pattern Φ', for any instance ϕ of Φ with its effective time interval $t_{\Phi\phi} \in T(\Phi)$, there exists an instance ϕ' of Φ' with its effective time interval $t_{\Phi'\phi'} \in T(\Phi')$ such that $t_{\Phi'\phi'} \subseteq t_{\Phi\phi}$, then we call Φ' a supporting sub-pattern of Φ.

For example, in Table 1, for the multi-temporal pattern $s_3 \overset{D}{\Rightarrow} s_1 \overset{O}{\Rightarrow} s_2, s_3 \overset{D}{\Rightarrow} s_1, s_1 \overset{O}{\Rightarrow} s_2$ and $s_3 \overset{O}{\Rightarrow} s_2$ are its sub-patterns, while both $s_3 \overset{D}{\Rightarrow} s_1$ and $s_1 \overset{O}{\Rightarrow} s_2$ are its sequential sub-patterns. Moreover, in Table 1, the multi-temporal pattern $\Phi := s_3 \overset{D}{\Rightarrow} s_1 \overset{O}{\Rightarrow} s_2$ has two instances $\phi_1 := e_{31} \overset{D}{\Rightarrow} e_{11} \overset{O}{\Rightarrow} e_{21}$ and $\phi_2 := e_{33} \overset{D}{\Rightarrow} e_{12} \overset{O}{\Rightarrow} e_{23}$ with effective time intervals $t_{\Phi\phi1} = (1, 8)$ and $t_{\Phi\phi2} = (28, 40)$, respectively. Now consider a sub-pattern $\Phi' := s_3 \overset{O}{\Rightarrow} s_2$, it has instances $\phi'_1 := e_{31} \overset{O}{\Rightarrow} e_{21}$ and $\phi'_2 := e_{33} \overset{O}{\Rightarrow} e_{23}$ with effective time intervals $t_{\Phi'\phi'1} = (1, 8)$ and $t_{\Phi'\phi'2} = (28, 38)$, respectively. Since $(1, 8) = t_{\Phi'\phi'1} \subseteq t_{\Phi\phi1} = (1, 8)$ and $(28, 38) = t_{\Phi'\phi'2} \subseteq t_{\Phi\phi2} = (28, 40)$, then Φ' is a supporting sub-pattern of Φ.

Property 2. *If Φ' is a supporting sub-pattern of Φ, then $supp(\Phi') \geqslant supp(\Phi)$.*

Proof. If $supp(\Phi)$ is the number of event instances that support Φ, it equals the number of effective time intervals. Since Φ' is a sub-pattern of Φ, so for an event instance, e.g., ϕ, supporting Φ, the effective time interval $t_{\Phi\phi}$ will cover a time interval $t_{\Phi'\phi'}$, where ϕ' supports Φ'. That is, $supp(\Phi') \geqslant supp(\Phi)$. □

For example, in Table 1, $supp(s_3 \overset{D}{\Rightarrow} s_1 \overset{O}{\Rightarrow} s_2) = 2$. For its supporting sub-pattern $s_3 \overset{O}{\Rightarrow} s_2$, we have $supp(s_3 \overset{O}{\Rightarrow} s_2) = 2 = supp(s_3 \overset{D}{\Rightarrow} s_1 \overset{O}{\Rightarrow} s_2)$. It could also be found that $s_1 \overset{O}{\Rightarrow} s_2$ is its supporting sub-pattern according to Definition 8. Furthermore, we can find $supp(s_1 \overset{O}{\Rightarrow} s_2) = 3 > 2 = supp(s_3 \overset{D}{\Rightarrow} s_1 \overset{O}{\Rightarrow} s_2)$.

Generally, Property 2 means that, for a temporal pattern, its *supp* value will not exceed its supporting sub-pattern's *supp* value. Then it could be inferred that, if a pattern is a frequent candidate pattern, any sub-pattern of the pattern is also a frequent candidate pattern, since $support(\Phi') = supp(\Phi')/|E| \geqslant supp(\Phi)/|E| = support(\Phi) \geqslant \alpha$. This property is quite important and could be used as a pruning strategy in further algorithmic design to improve algorithm's effectiveness. Additionally, due to the partial-ordering of temporal relationships, however, whether certain sub-patterns are supporting other sub-patterns needs to be further investigated.

Property 3. *If Φ' is a sequential sub-pattern of Φ, then Φ' is a supporting sub-pattern of Φ.*

Proof. For a sequential sub-pattern Φ' of Φ, there exists $t_{\Phi'\phi'} \in T(\Phi')$ such that $t_{\Phi'\phi'} \subseteq t_{\Phi\phi}, t_{\Phi\phi} \in T(\Phi)$, where ϕ and ϕ' are two event instances supporting Φ and Φ', respectively, and $t_{\Phi\phi}$ and $t_{\Phi'\phi'}$ are the corresponding effective time intervals of event instances ϕ and ϕ' for Φ and Φ', respectively. According to Definition 8, it can be derived that Φ' is a supporting sub-pattern of Φ. □

As an example in Table 1, both $s_3 \overset{D}{\Rightarrow} s_1$ and $s_1 \overset{O}{\Rightarrow} s_2$ are sequential sub-patterns of $s_3 \overset{D}{\Rightarrow} s_1 \overset{O}{\Rightarrow} s_2$, which means they are two supporting sub-patterns of $s_3 \overset{D}{\Rightarrow} s_1 \overset{O}{\Rightarrow} s_2$. Next, let us consider certain cases of non-sequential sub-patterns where temporal predicates D and B are of particular relevance and interest.

Property 4. *If an event instance $\phi := e_{0p0} \overset{R_1}{\Rightarrow} e_{1p1} \overset{R_2}{\Rightarrow} e_{2p2} \overset{R_3}{\Rightarrow} \ldots \overset{R_k}{\Rightarrow} e_{kpk}$ exists, $R_j \in \{D, B, O, E\}, j = 1, 2, \ldots, k$, then:*

(1) If $R_j = D, j = 1, 2, \ldots, k$, then $e_{ipi} \overset{R_k}{\Rightarrow} e_{kpk}$ exists, $i = 0, 1, 2, \ldots, k-1$.
(2) If $R_j = B, j = 1, 2, \ldots, k$, then $e_{(k-1)p(k-1)} \overset{R_k}{\Rightarrow} e_{kpk}$ exists. Moreover, if e_{kpk} happens in the *window* constraint of $e_{ipi}, i = 0, 1, 2, \ldots, k-1$, then $e_{ipi} \overset{R_k}{\Rightarrow} e_{kpk}$ exists, $i = 0, 1, \ldots, k-1$.

Proof

(1) If $R_j = D, j = 1, 2, \ldots, k$, then the effective time interval of event instance $\phi := e_{0p0} \overset{R_1}{\Rightarrow} e_{1p1} \overset{R_2}{\Rightarrow} e_{2p2} \overset{R_3}{\Rightarrow} \ldots \overset{R_k}{\Rightarrow} e_{kpk}$ can be denoted as (st_ϕ, et_ϕ). We have $st_{ipi} \leqslant st_\phi = st_{kpk} < et_{ipi} \leqslant et_{\Phi\phi} = et_{kpk}$ according to Theorem 1 ((1) and (3)). So the event instance $e_{ipi} \overset{R_k}{\Rightarrow} e_{kpk}$ exists, $i = 0, 1, 2, \ldots, k-1$.
(2) If $R_j = B, j = 1, 2, \ldots, k$, then the effective time interval of event instance $\phi' := e_{0p0} \overset{R_1}{\Rightarrow} e_{1p1} \overset{R_2}{\Rightarrow} e_{2p2} \overset{R_3}{\Rightarrow} \ldots \overset{R_k-1}{\Rightarrow} e_{(k-1)p(k-1)}$ can be denoted as $(st_{\phi'}, et_{\phi'})$. We have $et_{(k-1)p(k-1)} = et_{\phi'}$ and $et_{ipi} \leqslant et_{\phi'}, i = 0, 2, \ldots, k-1$, according to Theorem 1 (4). Based on the concept of *during* in Definition 2, we have $0 \leqslant st_{kpk} - et_{\phi'} \leqslant window$. That is, $e_{(k-1)p(k-1)} \overset{R_k}{\Rightarrow} e_{kpk}$ exists.

Moreover, it could be inferred that $st_{kpk} \geqslant et_{(k-1)p(k-1)} = et_{\phi'} \geqslant et_{ipi}$, if $st_{kpk} - et_{ipi} \leqslant window$, which means the *window* constraint is satisfied for events e_{kpk} and e_{ipi}, then $e_{ipi} \overset{R_k}{\Rightarrow} e_{kpk}$ exists, $i = 0, 1, \ldots, k-1$, otherwise not. □

It is worth mentioning that, if $R_k = O$ or E, event instance $e_{ipi} \overset{R_k}{\Rightarrow} e_{kp}$ may not happen. Subsequently, the following can be inferred.

Property 5. *Given a multi-temporal pattern* $\Phi := s_0 \overset{R_1}{\Rightarrow} s_1 \overset{R_2}{\Rightarrow} s_2 \overset{R_3}{\Rightarrow} \ldots \overset{R_k}{\Rightarrow} s_k, R_j \in \{D, B, O, E\}, j = 1, 2, \ldots, k$. *we have:*

(1) If $R_k = D$, the 1-item sub-pattern $s_i \overset{R_k}{\Rightarrow} s_k, i = 0, 1, 2, \ldots, k-1$, is a supporting sub-pattern of Φ.
(2) If $R_k = B$, the 1-item sub-pattern $s_{k-1} \overset{R_k}{\Rightarrow} s_k$ is a supporting sub-pattern of Φ. Moreover, if s_i and s_k satisfy the *window* constrain, then the 1-item sub-pattern $s_i \overset{R_k}{\Rightarrow} s_k$ is a supporting sub-pattern of $\Phi, i = 0, 2, \ldots, k-1$.

Proof. The property can be proved directly based on Property 4. □

Property 6. *Given a multi-temporal pattern* $\Phi := s_0 \overset{R_1}{\Rightarrow} s_1 \overset{R_2}{\Rightarrow} s_2 \overset{R_3}{\Rightarrow} \ldots \overset{R_k}{\Rightarrow} s_k, R_j \in \{D, B, O, E\}, j = 1, 2, \ldots, k$, *if* $R_k = D$, *the sub-pattern* $s_0 \overset{R_1}{\Rightarrow} s_1 \overset{R_2}{\Rightarrow} s_2 \overset{R_3}{\Rightarrow} \ldots \overset{R_{k-2}}{\Rightarrow} s_{k-2} \overset{R_k}{\Rightarrow} s_k$ *is a supporting sub-pattern of* Φ.

Proof. Without loss of generality, we denote $s_0 \overset{R_1}{\Rightarrow} s_1 \overset{R_2}{\Rightarrow} s_2 \overset{R_3}{\Rightarrow} \ldots \overset{R_{k-2}}{\Rightarrow} s_{k-2}$ as Φ'. Then Φ can be denoted as $\Phi' \overset{R_{k-1}}{\Rightarrow} s_{k-1} \overset{R_k}{\Rightarrow} s_k$. According to Property 5, it could be inferred that $\Phi' \overset{R_k}{\Rightarrow} s_k$ is a supporting sub-pattern of Φ. Thus, the sub-pattern $s_0 \overset{R_1}{\Rightarrow} s_1 \overset{R_2}{\Rightarrow} s_2 \overset{R_3}{\Rightarrow} \ldots \overset{R_{k-2}}{\Rightarrow} s_{k-2} \overset{R_k}{\Rightarrow} s_k$ is a supporting sub-pattern of Φ. □

Property 7. *Given a multi-temporal pattern* $\Phi := s_0 \overset{R_1}{\Rightarrow} s_1 \overset{R_2}{\Rightarrow} s_2 \overset{R_3}{\Rightarrow} \ldots \overset{R_k}{\Rightarrow} s_k, R_j \in \{D, B, O, E\}, j = 1, 2, \ldots, k$, *if there exists any* $R_j = D$ *or* $B, j = 1, 2, \ldots, k-1$, *then* $s_1 \overset{R_2}{\Rightarrow} s_2 \overset{R_3}{\Rightarrow} \ldots \overset{R_k}{\Rightarrow} s_k$ *is a supporting sub-pattern of* Φ; *Otherwise, only if* $R_k = D$ *or* B, *then* $s_1 \overset{R_2}{\Rightarrow} s_2 \overset{R_3}{\Rightarrow} \ldots \overset{R_k}{\Rightarrow} s_k$ *is a supporting sub-pattern of* Φ.

Proof. It can be proved directly based on Property 6. □

Properties 4–7 reflect the characteristics of temporal predicates D and B for non-sequential sub-patterns, which can be further integrated into the mining algorithm along with other properties. For example, in Table 1, instance $e_{61} \overset{D}{\Rightarrow} e_{21} \overset{D}{\Rightarrow} e_{51}$ exists, then $e_{61} \overset{D}{\Rightarrow} e_{51}$ exists according to Property 4. A multi-temporal pattern $s_6 \overset{O}{\Rightarrow} s_3 \overset{D}{\Rightarrow} s_1$ exists in Table 1, then both $s_3 \overset{D}{\Rightarrow} s_1$ and $s_6 \overset{D}{\Rightarrow} s_1$ are supporting sub-patterns of $s_6 \overset{O}{\Rightarrow} s_3 \overset{D}{\Rightarrow} s_1$ according to Property 5. A multi-temporal pattern $s_4 \overset{O}{\Rightarrow} s_6 \overset{D}{\Rightarrow} s_2 \overset{D}{\Rightarrow} s_5$ exists in Table 1, then $s_4 \overset{O}{\Rightarrow} s_6 \overset{D}{\Rightarrow} s_5$ is a supporting sub-pattern of $s_4 \overset{O}{\Rightarrow} s_6 \overset{D}{\Rightarrow} s_2 \overset{D}{\Rightarrow} s_5$ according to Property 6, and $s_6 \overset{D}{\Rightarrow} s_2 \overset{D}{\Rightarrow} s_5$ is a supporting sub-pattern of $s_4 \overset{O}{\Rightarrow} s_6 \overset{D}{\Rightarrow} s_2 \overset{D}{\Rightarrow} s_5$ according to Property 7. For temporal predicate B, Properties 4–7 could be exemplified in a similar manner.

4. The mining algorithm

4.1. Generating 2-item temporal patterns (**k = 1**)

In order to facilitate the mining process and the corresponding algorithm, the temporal database needs to be re-organized. First, a new table should be constructed with attributes (fields) being states. Second, all the temporal events should be assigned into corresponding columns. For example, Table 1 could be transformed into Table 2.

Only the events that occur in an acceptable length of time, satisfying the *window* constraint, will be regarded to have a temporal relationship. This means that, if two events are apart from each other far beyond the realm of the window of interest, the two events are not regarded to have a temporal relationship that is worth being considered.

Given a temporal database D_T with the set of states S, initially, a set of degenerated $(k = 0)$ patterns could be constructed, e.g., $C_0 = \{s_j | s_j \in S\}$. Then, the temporal database could be scanned to determine whether each s_j is frequent or not. Hence a set of frequent degenerated 1-item $(k = 0)$ patterns could be obtained, e.g., $F_0 = \{s_j | support(s_j) \geqslant \alpha, s_j \in S\}$. Suppose there are

Table 2
Transformed temporal database D'_T.

$T(s_1)$	$T(s_2)$	$T(s_3)$	$T(s_4)$	$T(s_5)$	$T(s_6)$	...
(1, 5)	(2, 8)	(1, 4)	(3, 7)	(2, 10)	(2, 5)	...
(30, 40)	(4, 6)	(33, 38)	(25, 27)	(27, 42)	(34, 39)	...
...	(28, 32)	(30, 38)	...	...	...	...
...	...	...	...	...	...	...

n events in the temporal database, the computational complexity of this process is at $O(n)$, i.e., one database scan to calculate the *supp* values of all states.

Moreover, based on the following property, 2-item $(k = 1)$ frequent patterns could be generated based on 1-item patterns.

Property 8. *For a 2-item temporal pattern Φ, e.g., $\Phi := s_i \overset{R}{\Rightarrow} s_j, R = \{D, B, O, E\}$, $supp(\Phi) \leqslant supp(s_j)$.*

Proof. Since s_j is a degenerated pattern, it is a sequential sub-pattern of Φ. According to Property 3, s_j is a supporting sub-pattern of Φ. Furthermore, $supp(\Phi) \leqslant supp(s_j)$ (Property 2). □

Upon Property 8, the set of 2-item $(k = 1)$ candidate patterns could be generated, e.g., $C_1 = \{s_i \overset{R}{\Rightarrow} s_j / \ s_i \in S, s_j \in F_0\}$. In this way, the efficiency will be improved, since there is no need for its complementary set $C_1' = \{s_i \overset{R}{\Rightarrow} s_j | s_i \in S, s_j \notin F_0\}$ to be generated or further calculated.

Thereafter, the temporal database could be scanned to get the set of 2-item frequent patterns, e.g., $F_1 = \{s_i \overset{R}{\Rightarrow} s_j | s_i \in S, s_j \in F_0, support(s_i \overset{R}{\Rightarrow} s_j) \geqslant \alpha\}$.

Moreover, the set of $(k+1)$-item candidate patterns, e.g., C_k, could be generated based on the set of k-item frequent patterns, e.g., F_{k-1}. Next, some further optimization could be made in order to improve the effectiveness of the mining process.

4.2. Generating (k + 1)-item temporal patterns ($k \geqslant 1$)

Based on previous discussions, the following property could be obtained.

Property 9. *Suppose $R_j \in \{D, B, O, E\}, j = 1, 2, \ldots, k, k \geqslant 2$, given a $(k-1)$-item temporal pattern Φ, $\Phi := s_0 \overset{R_1}{\Rightarrow} s_1 \overset{R_2}{\Rightarrow} s_2 \overset{R_3}{\Rightarrow} \ldots \overset{R_k}{\Rightarrow} s_k$, and three sub-patterns of Φ, e.g., Φ_1, Φ_2 and Φ_3, e.g., $\Phi_1 := s_0 \overset{R_1}{\Rightarrow} s_1 \overset{R_2}{\Rightarrow} s_2 \overset{R_3}{\Rightarrow} \ldots \overset{R_{k-1}}{\Rightarrow} s_{k-1}$, $\Phi_2 := s_1 \overset{R_2}{\Rightarrow} s_2 \overset{R_3}{\Rightarrow} \ldots \overset{R_k}{\Rightarrow} s_k$, and $\Phi_3 := s_0 \overset{R_1}{\Rightarrow} s_1 \overset{R_2}{\Rightarrow} s_2 \overset{R_3}{\Rightarrow} \ldots \overset{R_{k-2}}{\Rightarrow} s_{k-2} \overset{R_k}{\Rightarrow} s_k$. We have:*

(1) If $R_k = D$, if either of Φ_1, Φ_2 and Φ_3 is not frequent, e.g., support $(X) < \alpha, X = \Phi_1, \Phi_2$ or Φ_3, then Φ is not frequent, i.e., support $\Phi) < \alpha$.
(2) If there exist any $R_j = D$ or $E, j = 1, 2, \ldots, k-1$, and either Φ_1 or Φ_2 is not frequent, then Φ is not frequent, i.e., support $(\Phi) < \alpha$.
(3) If $R_j \neq D$ and $R_j \neq E$, for $\forall j = 1, 2, \ldots, k-1$, and $R_k = B$, and either Φ_1 or Φ_2 is not frequent, then Φ is not frequent, i.e., support $(\Phi) < \alpha$.

Proof

(1) According to Property 3, Φ_1 is a supporting sub-pattern of Φ. According to Property 5, Φ_2 is a supporting sub-pattern of Φ. According to Property 6, Φ_3 is a supporting sub-pattern of Φ. So we have $supp(\Phi_1) \geqslant supp(\Phi), supp(\Phi_2) \geqslant supp(\Phi)$, and $supp(\Phi_3) \geqslant supp(\Phi)$. That means that if either of Φ_1, Φ_2, and Φ_3 is not frequent, then Φ is not frequent.
(2) This can be proved in the same way based on Properties 5 and 6.
(3) Likewise, this can also be proved in the same way based on Properties 5 and 6. □

In generating C_k based on F_{k-1}, Property 9 could be used to filter the non-frequent patterns without scanning the database. The procedure is as follows:

(1) For each k-item frequent pattern, e.g., $\Phi_1 := s_0 \overset{R_1}{\Rightarrow} s_1 \overset{R_2}{\Rightarrow} s_2 \overset{R_3}{\Rightarrow} \ldots \overset{R_{k-1}}{\Rightarrow} s_{k-1}$, in F_{k-1}, label it whether the pattern contains D or E.
(2) Search in F_{k-1} to get $\Phi_2 := s_1 \overset{R_2}{\Rightarrow} s_2 \overset{R_3}{\Rightarrow} \ldots \overset{R_k}{\Rightarrow} s_k$, where R_k and s_k are new temporal predicate and new state corresponding to Φ_1. We call Φ_2 matches Φ_1. Then a new $(k+1)$-item candidate pattern Φ, e.g., $\Phi := s_0 \overset{R_1}{\Rightarrow} s_1 \overset{R_2}{\Rightarrow} s_2 \overset{R_3}{\Rightarrow} \ldots \overset{R_k}{\Rightarrow} s_k$, can be generated. Thus, the set of $(k+1)$-item candidate patterns, denoted as C_k, can be constructed, e.g., $C_k = \{\Phi | \Phi := s_0 \overset{R_1}{\Rightarrow} s_1 \overset{R_2}{\Rightarrow} s_2 \overset{R_3}{\Rightarrow} \ldots \overset{R_k}{\Rightarrow} s_k$, where $\Phi_1 := s_0 \overset{R_1}{\Rightarrow} s_1 \overset{R_2}{\Rightarrow} s_2 \overset{R_3}{\Rightarrow} \ldots \overset{R_{k-1}}{\Rightarrow} s_{k-1}, \Phi := s_0 \overset{R_1}{\Rightarrow} s_1 \overset{R_2}{\Rightarrow} s_2 \overset{R_3}{\Rightarrow} \ldots \overset{R_k}{\Rightarrow} s_k$, and $\Phi_1, \Phi_2 \in F_{k-1}\}$.
(3) Based on generated C_k, before scanning the database to calculate the degrees of support and confidence for each $\Phi, \Phi \in C_k$, to determine whether each Φ is a qualified multi-temporal pattern, some filtering strategies could be formed:
   (a) If $R_k = D$, it means that Φ_3, e.g., $\Phi_3 := s_0 \overset{R_1}{\Rightarrow} s_1 \overset{R_2}{\Rightarrow} s_2 \overset{R_3}{\Rightarrow} \ldots \overset{R_{k-2}}{\Rightarrow} s_{k-2} \overset{R_k}{\Rightarrow} s_k$, corresponding to Φ should also be in F_{k-1}. Otherwise, Φ will not be a qualified pattern (Property 9 (1)). Then scan F_{k-1} to determine whether $\Phi_3 \in F_{k-1}$. If it is not, then delete it from C_k, i.e., $C_k = C_k - \{\Phi_3\}$.
   (b) If $R_k \neq D$ and there exists any $R_j = D$ or $E, j = 1, 2, \ldots, k-1$, then keep C_k unchanged (Property 9 (2)).
   (c) If $R_j \neq D$ and $R_j \neq E$, for $\forall j = 1, 2, \ldots, k-1$, and $R_k = B$, then keep C_k unchanged (Property 9 (2)).

Table 3
Algorithm for generating frequent temporal patterns.

1	$S = \{s \mid s \text{ is a state in the } D_T\}$;
2	$C_0 = \{s_i \mid s_i \in S\}$
3	$F_0 = \emptyset$;
4	$m = \lvert S \rvert$;
5	FOR $s_j \in S$ DO
6	$F_0 = F_0 \cup \{s_j \mid support(s_j) \geqslant \alpha, s_j \in S\}$
7	ENDFOR
8	$C_1 = \{s_i \overset{R}{\Rightarrow} s_j \mid s_i \in S, s_j \in F_0\}$;
9	$k = 1$;
10	WHILE $C_k \neq \emptyset$ DO
11	$F_k = \emptyset$;
12	FOR $\Phi_k \in C_k$ DO
13	Calculating $support(\Phi_k)$;
14	IF $support(\Phi_k) \geqslant \alpha$ THEN
15	$F_k = F_k \cup \{\Phi_k\}$;
16	END IF
17	END FOR
18	$C_{k+1} = \emptyset$;
19	FOR $\Phi_1 \in F_k$ DO
20	IF Φ_1 contains D or E THEN
21	Flag=1;
22	END IF
23	FOR $\Phi_2 \in F_k$ DO
24	IF Φ_2 matches Φ_1 AND $\Phi_3 \in F_k$ THEN
25	$C_{k+1} = C_{k+1} \cup \Phi$;
26	END IF
27	Filter C_{k+1};
28	END FOR
29	END FOR
30	END WHILE

Table 4
Algorithm for generating qualified temporal patterns.

1	$Q = \emptyset$;
2	$k = 0$;
3	WHILE $F_k \neq \emptyset$ DO
4	FOR $\Phi \in F_k$
5	Calculating $confidence(\Phi)$;
6	IF $confidence(\Phi) \geqslant \beta$ THEN
7	$Q = Q \cup \{\Phi\}$;
8	END IF
9	END FOR
10	$k = k + 1$;
11	END WHILE

Then, based on the updated C_k, scan the database to calculate the *supp* value of each (k+1)-item candidate Φ in C_k. Finally, F_k could be obtained. Moreover, based on F_k, C_{k+1} could be generated and filtered similarly. When $C_k = \emptyset$, then terminate, which means that all the frequent patterns have been discovered.

The algorithmic detail for generating all frequent temporal patterns is shown in Table 3.

Furthermore, based on the generated frequent temporal patterns along with the *supp* values, the degree of confidence for each temporal pattern could be calculated. Finally, the set of qualified temporal patterns could be obtained. The algorithm for calculating the degrees of confidence to generate qualified patterns is shown as follows (Table 4).

5. An application of the proposed approach to associative movement of stock markets

Associative movement of stock markets is an important issue in financial studies [16,17,27]. With the fast development of China's economy, the stock markets in China (e.g., Shanghai stock market, Shenzhen stock market and Hong Kong stock market) play an important role in China's economy, and are also attracting attention from the world economy perspective. Due to the historical reasons, Hong Kong's economy was to a certain extent independent of Chinese mainland economy for decades. However, in the context of globalization and the return of Hong Kong, the associative movement of stock markets between Chinese mainland and Hong Kong was deemed noteworthy in recent years, giving rise to several research efforts [22,28,30].

Table 5
Description of data.

Level	Data description
Industry[a]	*Major industry indexes of Chinese mainland market* CSI 300 Energy Index (CSI 300 EI) CSI 300 Materials Index (CSI 300 MI) CSI 300 Industrials Index (CSI 300 II) CSI 300 Consumer Index (CSI 300 CI) CSI 300 Telecommunication Services Index (CSI 300 TS) CSI 300 Utilities Index (CSI 300 UI) CSI 300 Financials Index (CSI 300 FI) CSI 300 Information Technology Index (CSI 300 IT) *Major industry indexes of and Hong Kong market* Hang Seng Energy Index (HS EI) Hang Seng Materials Index (HS MI) Hang Seng Industrials Index (HS II) Hang Seng Consumer Index (HS CI) Hang Seng Telecommunication index (HS TI) Hang Seng Utilities Index (HS UI) Hang Seng Financials Index (HS FI) Hang Seng Information Technology Index (HS IT)
Market[b]	*Major composite indexes of Chinese mainland market and Hong Kong market:* Hang Seng China-Affiliated Corporations Index (HSCACI) Hang Seng China Enterprises Index (HSCEI) Hang Seng Freefloat Composite Index (HSFCI) Shanghai Composite Index (SCI) Shenzhen Component Index (SCI2)

[a] Two indexes in CSI 300 Industry Indexes (CSI 300 Consumer Discretionary Index and CSI 300 Health Care Index) and three indexes in Hang Seng Industry Indexes (Hang Seng Services Index, Hang Seng Construction Index and Hang Seng Composite Index) are omitted, since each of them cannot be found a corresponding index in the other market.
[b] These five indexes are the major market indexes for Chinese mainland market and Hong Kong market.

Concretely, two research questions are of interest: one is whether or not these two markets have significant associative movements; the other is what will the direction of the movement be if there is any significant associative movement.

From the viewpoint of temporal patterns, associative movement corresponds to temporal relationships. In this section, we will use the proposed approach to analyze the associative movement of Chinese mainland stock market and Hong Kong stock market. In doing so, first, the data will be described, along with the possible pre-processing procedure. Next, the associative movements from two levels, namely, the industry-level and market-level, will be analyzed in light of respective indexes.

5.1. Data description and pre-processing

All the raw data, including daily stock price data, industry analysis and index data, were from Wind finance database [33] for the period of July 2003 to January 2008 (excluding January–June 2003 while SARS was severely epidemic). The description of data is shown in Table 5.

Since the original data were listed with the values of every day, they need to be pre-processed before being used in mining multi-temporal patterns. First, in order to reduce the noises, we used 5-days moving average values of the index values. Second, without loss of generality, we consider two states of each index: *increase* and *decrease* (Totally, for K indexes, there will be $2 * K$ states). Simply, if the value in day $i + 1$ is higher than the value in day i, then this case is referred to as the event with a state being *increase* starting at i and ending at $i + 1$. Since the value of an index could keep *increase*/*decrease* for multiple days, we merged the continuous events of *increase* or *decrease* into one event, e.g.,{index A is *decrease*, 6, 8} representing that index A's value would keep *decrease* from day 6 to day 8. Third, after the transforming and merging, the events of all the indexes could be integrated into a temporal database.

Fig. 1 exemplifies the pre-processing procedure on some real data of Shanghai Composite Index and Shenzhen Composite Index.

5.2. Discovered multi-temporal patterns

After some testing, based on the pre-processed temporal data, set $\alpha = 0.25, \beta = 0.3$, the discovered industry-level multi-temporal patterns are as shown in Table 6.

From the discovered industry-level multi-temporal patterns, some remarks could be made.

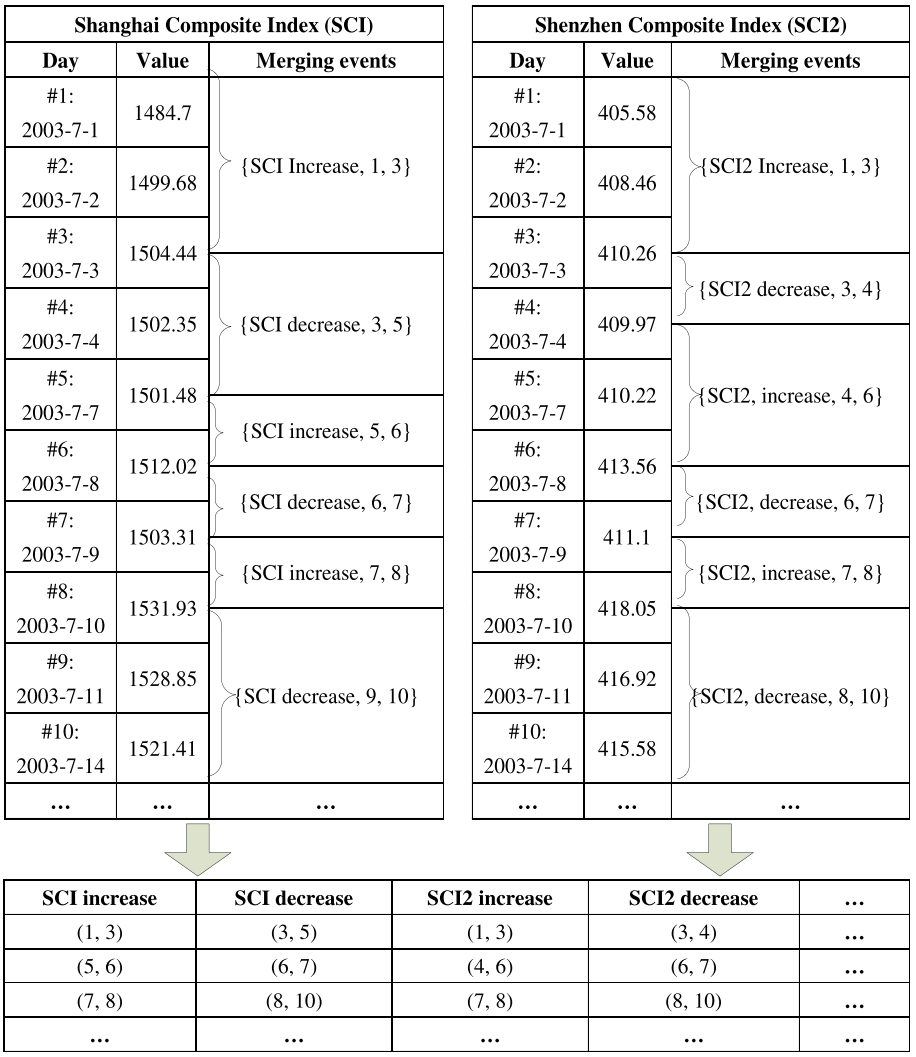

Fig. 1. Data pre-processing.

First, Hang Seng industry-level indexes had information advantage over CSI 300 industry-level indexes (#1–#7,#13–#14), especially on Telecommunications, Energy, Financials, Industrials, IT and Utilities, which still reveals a fact that Hong Kong market was more globalized and sensitive to news and information, while Chinese mainland market was not so sensitive compared with Hong Kong market.

Second, the Materials industry of Chinese mainland market had information advantage over the Materials and Industrials industries of Hong Kong market (#8–#12), since Chinese mainland market was one of the major markets of materials, while Hong Kong market was not and the Materials and Industrials industries in Hong Kong mainly purchased materials from Chinese mainland market. Moreover, as the precedent industries of Materials, Energy and Consumer had information advantage over Materials in Chinese mainland market (#9–#12), since they were the major industries that require materials.

Third, the Financials industry had information advantage over Telecommunications industry in Hong Kong market (#13–#14), since Hong Kong was one of the financial centers in the world and the Financials industry was more sensitive to news and information.

Overall, the results show that the associative movement of Chinese mainland market and Hong Kong market was quite significant on the industry-level.

5.3. Discovered market-level temporal patterns

Also after several rounds of testing, based on the pre-processed temporal data, set $\alpha = 0.3, \beta = 0.4$. On the market-level temporal database, the discovered multi-temporal patterns contain predicates *before* and *during*, and the states in a discov-

Table 6
Discovered industry-level multi-temporal patterns.

#	*Discovered industry-level multi-temporal patterns*
1	Hang Seng Telecommunications Index *increase* $\overset{B}{\Rightarrow}$CSI 300 Telecommunications Index *increase*
2	Hang Seng Energy Index *increase* $\overset{B}{\Rightarrow}$CSI 300 Energy Index *increase*
3	Hang Seng Financials Index *increase* $\overset{B}{\Rightarrow}$CSI 300 Energy Index *increase*
4	Hang Seng Industrials Index *increase* $\overset{B}{\Rightarrow}$CSI 300 Consumer Index *increase*
5	Hang Seng IT Index *increase* $\overset{B}{\Rightarrow}$CSI 300 Telecommunication Index *increase*
6	Hang Seng IT Index *increase* $\overset{B}{\Rightarrow}$CSI 300 IT Index *increase*
7	Hang Seng Utilities Index *increase* $\overset{B}{\Rightarrow}$CSI 300 Utilities Index *increase*
8	CSI 300 Materials Index *increase* $\overset{B}{\Rightarrow}$ Hang Seng Materials Index *increase*
9	CSI 300 Energy Index *increase* $\overset{B}{\Rightarrow}$CSI 300 Materials Index *increase* $\overset{B}{\Rightarrow}$Hang Seng Materials Index *increase*
10	CSI 300 Consumer Index *increase* $\overset{B}{\Rightarrow}$CSI 300 Materials Index *increase* $\overset{B}{\Rightarrow}$Hang Seng Industrials Index *increase*
11	CSI 300 Energy Index *increase* $\overset{B}{\Rightarrow}$CSI 300 Materials Index *increase* $\overset{B}{\Rightarrow}$Hang Seng Industrials Index *increase*
12	CSI 300 Consumer Index*increase* $\overset{B}{\Rightarrow}$CSI 300 Materials Index *increase* $\overset{B}{\Rightarrow}$Hang Seng Materials Index *increase*
13	Hang Seng Financials Index *increase* $\overset{B}{\Rightarrow}$Hang Seng Telecommunications Index *increase* $\overset{B}{\Rightarrow}$CSI 300 Industrials Index *increase*
14	Hang Seng Financials Index *increase* $\overset{B}{\Rightarrow}$Hang Seng Telecommunications Index *increase* $\overset{B}{\Rightarrow}$CSI 300 Materials Index *increase*

Table 7
Discovered 2-item market-level temporal patterns.

P ⇒ C	HSCACI	HSCEI	HSFCI	SCI	SCI2
HSCACI	N/A	B	N/A	B	B
HSCEI	B/D	N/A	B	B/D	N/A
HSFCI	D	D	N/A	B	N/A
SCI	B/D	D	B	N/A	N/A
SCI2	N/A	N/A	N/A	B/D	N/A

Table 8
Discovered 3-item market-level temporal patterns.

#	Discovered 3-item market-level multi-temporal patterns
1	HSFCI *decrease* $\overset{D}{\Rightarrow}$ HSCACI *decrease* $\overset{B}{\Rightarrow}$ SCI *decrease*
2	HSFCI *increase* $\overset{D}{\Rightarrow}$ HSCACI *increase* $\overset{D}{\Rightarrow}$ HSCEI *increase*

ered multi-temporal pattern are the same, e.g., HSFCI *decrease* $\overset{D}{\Rightarrow}$HSCACI *decrease* $\overset{B}{\Rightarrow}$SCI *decrease*. For simplicity, we omit states *increase* and *decrease* in the representation as shown. Since the 2-item temporal patterns are so frequent, for clarity, we represent the 2-item temporal patterns in Table 7. Each item in the first column is the precedence of a 2-item pattern, while the item in the first row is the consequence of a 2-item pattern, the symbol "B" or "D" in the table shows the corresponding predicate "*before*" or "*during*", and the symbol "B/D" shows that the corresponding two items have temporal relationships "*before*" and "*during*" simultaneously. The symbol "N/A" represents that there is no significant temporal relationship between the corresponding two items.

For k-item temporal patterns ($k > 2$), there are two 3-item patterns discovered as shown in Table 8.

Note that pattern 2 in Table 8 (i.e., HSFCI *decrease* $\overset{D}{\Rightarrow}$ HSCACI *decrease* $\overset{B}{\Rightarrow}$ SCI *decrease*) is a multi-temporal pattern with *During* and *Before*, which was generated from its frequent sub-sequential patterns such as "HSFCI *decrease*", "HSCACI *decrease*", "SCI *decrease*", "HSFCI *decrease* $\overset{D}{\Rightarrow}$HSCACI *decrease*" and "HSCACI *decrease* $\overset{B}{\Rightarrow}$SCI *decrease*", according to Properties 2, 3 and 9.

Moreover, from Tables 7 and 8, we could have the following findings. First, HSCACI had a certain information advantage over other market indexes. Second, HSCEI, HSFCI and SCI were closely related and none of these three indexes had significant information advantages over others. Third, SCI2 was closely related to SCI but the temporal relationships between SCI2 and indexes in Hong Kong market were not significant, which reveals that, compared with Shanghai market, Shenzhen market was less related to Hong Kong market. Overall, the results show that the associative movement of Chinese mainland market and Hong Kong market was quite significant on the market-level.

6. Conclusion and future work

In this paper, the notion of multi-temporal patterns with four temporal predicates (i.e., *Before, During, Overlap* and *Equal*) has been presented as important forms of knowledge for discovery. In-depth investigations of several properties relating to the combinations of temporal predicates in patterns, sequential/non-sequential sub-patterns, and the pattern generation have been conducted so as to develop effective optimization strategies for reducing the database scan in the generation of candidate patterns of the mining process. Correspondingly, the proposed approach has also provided algorithmic details. Finally, the approach has then been applied to stock markets to explore possible associative movements between the stock markets of Chinese mainland and Hong Kong, revealing that there were significant associative movements between the two markets, at both industry and market-levels.

Future work could be centered on further theoretical scalability analyses along with other synthetic and real data experiments, on extensions of the approaches to more temporal relationships, and on other financial applications.

Acknowledgements

The work was partly supported by the National Natural Science Foundation of China (70890083/70621061), the MOE Project of Key Research Institute of Humanities and Social Sciences at Universities of China (07JJD63005) and Tsinghua University's Research Center for Contemporary Management.

References

[1] R. Agrawal, R. Srikant, Fast algorithms for mining association rules in large databases, in: Proceedings of the 20th International Conference on Very Large Data Bases, Santiago, Chile, September 12–15, 1994, pp. 487–499.
[2] R. Agrawal, R. Srikant, Mining sequential patterns, in: Proceedings of the Eleventh International Conference on Data Engineering, Taiwan, March 06–10, 1995, pp. 3–14.
[3] R. Agrawal, T. Imielinski, A. Swami, Mining association rules between sets of items in large databases, in: Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, vol. 10, Washington, DC, May 26–28, 1993, pp. 207–216, .
[4] J.F. Allen, Maintaining knowledge about temporal intervals, Communications of the ACM 26 (11) (1983) 832–P843.
[5] E. Chen, H. Cao, Q. Li, T. Qian, Efficient strategies for tough aggregate constraint-based sequential pattern mining, Information Sciences 178 (6) (2008) 1498–1518.
[6] G. Chen, J. Ai, W. Yu, Discovering temporal association rules for time-lag data, in: Proceedings of International Conference on E-business (ICEB2002), Beijing, May 2002, pp. 324–328.
[7] Y.-L. Chen, Y.-H. Hu, Constraint-based sequential pattern mining: the consideration of recency and compactness, Decision Support Systems 42 (2) (2006) 1203–1215.
[8] Y.-L. Chen, S.-Y. Wu, Mining temporal patterns from sequence database of interval-based events, in: FSKD 2006, LNAI 4223, 2006, pp. 586–595.
[9] G. Das, D. Gunopulos, H. Mannila, Finding similar time series, in: J. Komorowski et al. (Eds), Proceedings of the First European Symposium on Principles of Data Mining and Knowledge Discovery, LNAI 1263, Springer, Trondheim, Norway, 1997, pp. 88–100.
[10] G. Das, K.I. Lin, H. Mannila, Rule discovery from time series, in: Proceedings of the Third International Conference on Knowledge Discovery and Data Mining, 1998, pp. 16–22.
[11] S. De Amo, D.A. Furtado, First-order temporal pattern mining with regular expression constraints, Data and Knowledge Engineering 62 (3) (2007) 401–420.
[12] F. Giannotti, M. Nanni, F. Pinelli, D. Pedreschi, Trajectory pattern mining, in: International Conference on Knowledge Discovery and Data Mining, Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining Table of Contents, San Jose, California, USA, 2007, pp. 330–339.
[13] F. Hoppner, Discovery of temporal patterns – learning rules about the qualitative behaviour of time series, in: PKDD'01, No. 2168 of LNAI, Freiburg, Germany, 2001, pp. 192–203.
[14] T. Hu, S.Y. Sung, H. Xiong, Q. Fu, Discovery of maximum length frequent itemsets, Information Sciences 178 (1) (2008) 69–87.
[15] Y.-C. Hu, G.-H. Tzeng, C.-M. Chen, Deriving two-stage learning sequences from knowledge in fuzzy sequential pattern mining, Information Sciences 159 (1–2) (2004) 69–86.
[16] B-N. Huang, C-W. Yang, W-S. Hu, Causality and cointegration of stock markets among the United States, Japan and South China growth triangle, International Review of Financial Analysis 9 (3) (2000) 281–297.
[17] B.N. Joen, G.M. Von Furstenberg, Growing international co-movement in stock price indexes, Quarterly Review of Economics and Finance 30 (1990) 15–30.
[18] P. Kam, A.W. Fu, Discovering Temporal Patterns for Interval-based Events Data Warehousing and Knowledge Discovery, Springer, Berlin/Heidelberg, 2000. pp. 317–236.
[19] A.J.T. Lee, C.-S. Wang, An efficient algorithm for mining frequent inter-transaction patterns, Information Sciences 177 (17) (2007) 3453–3476.
[20] J.T. Lee, Y. Chen, W-C. Ip, Mining frequent trajectory patterns in spatial–temporal databases, Information Sciences 179 (13) (2009) 2218–2231.
[21] J.T. Lee, C.-S. Wang, An efficient algorithm for mining frequent inter-transaction patterns, Information Sciences 177 (17) (2007) 3453–3476.
[22] Y. Li, J.F. Greco, B. Chavis, Lead–lag relations between A shares and H shares in the Chinese stock markets, Working Paper, California State University, Fullerton, 2000.
[23] Y. Li, S. Zhu, X.S. Wang, S. Jajoodia, Looking into the seeds of time: discovering temporal patterns in large transaction sets, Information Sciences 176 (8) (2006) 1003–1031.
[24] M.-Y. Lin, S.-Y. Lee, Interactive sequence discovery by incremental mining, Information Sciences 165 (3–4) (2004) 187–205.
[25] M-Y. Lin, S-C. Hsueh, C-W. Chang, Fast discovering of sequential patterns in large databases using effective time-indexing, Information Sciences 178 (22) (2008) 4228–4245.
[26] M.-Y. Lin, S.-C. Hsueh, C.-W. Chang, Fast discovery of sequential patterns in large databases using effective time-indexing, Information Sciences 178 (22) (2008) 4228–4245.
[27] D.B. Panton, V.P. Lessig, O.M. Joy, Comovement of international equity markets: a taxonomic approach, The Journal of Financial and Quantitative Analysis 11 (3) (1976) 415–432.
[28] W.P.H. Poon, H-G. Fund, Red chips or H-shares: which China-backed securities process information the fastest, Journal of Multinational Financial Management 10 (2000) 315–343.
[29] C.P. Rainsford, J.F. Roddic, Adding temporal semantics to association rules, in: Proceedings of the Third European Conference on Principles and Practice of Knowledge Discovery in Databases, PKDD '99, Prague, Czech Republic. September 15–18, 1999, pp. 504–509.

[30] Q. Sun, W.H.S. Tong, The effect of market segmentation on stock prices: the China syndrome, Journal of Banking and Finance 24 (2000) 1875–1902.
[31] T-J. Tsay, T-J. Hsu, J-R. Yu, FIUT: a new method for mining frequent itemsets, Information Sciences 179 (11) (2009) 1724–1737.
[32] E. Winarko, J.F. Roddick, ARMADA – an algorithm for discovering richer relative temporal association rules from interval-based data, Data and Knowledge Engineering 63 (1) (2007) 76–90.
[33] Wind Financial Databases, Shanghai Wind Information Co., Ltd., http://www.wind.com.cn/.
[34] S-Y. Wu, Y-L. Chen, Mining nonambiguous temporal patterns for interval-based events, IEEE Transactions on Knowledge and Data Engineering 19 (6) (2007) 742–758.
[35] W. Yu, G. Chen, Mining delayed association rules based on temporal data, Computer Application and Research 12 (2002) 19–22 (in Chinese).
[36] M.J. Zaki, SPADE: an efficient algorithm for mining frequent sequences, Machine Learning 42 (1/2) (2000) 31–60.
[37] L. Zhang, G. Chen, T. Brijs, X. Zhang, Discovering during-temporal patterns (DTPs) in large temporal databases, Expert Systems with Applications 34 (2) (2008) 1178–1189.