# TEMPORAL ASSOCIATION RULE MINING BASED ON T-APRIORI ALGORITHM AND ITS TYPICAL APPLICATION

ZHAI Liang[a, b *], TANG Xinming[b, c], LI Lin[a], JIANG Wenliang[a]

[a] School of Resource and Environment Science, Wuhan University, 129 Luoyu Road, Wuhan, 430079, China - : zhailiang@126.com, jwl22@163.com, lilin@telecarto.com
[b] Key Laboratory of Geo-informatics of State Bureau of Surveying and Mapping, Chinese Academy of Surveying and Mapping, 16 Beitaiping Road, Haidian District, Beijing, 100039, China - : tang@casm.ac.cn
[c] Institute of Remote Sensing and Geographic Information Systems, Peking University, Beijing, 100871, China

**KEYWARDS:** Data mining, Temporal Association Rule Mining, A Sequence of Ecological Events

**ABSTRACT:**

A new algorithm "T-Apriori" based on time constraint is designed and implemented on the basis of analyzing the related definitions and general steps of temporal association rule mining. The concepts of ecological event and sequence of ecological events are proposed and the problems of temporal association rule mining based on a sequence of ecological events are discussed in details. A sequence of ecological events is an ordered ecological events set. Red tide phenomena of noctiluca scintillans during 1991 and 1992 in Dapeng bay, South China Sea was taken as an example to validate T-Apriori algorithm and K-means clustering analysis is used to map the quantitative association rule problem into the Boolean association rule problem. Our example experiment shows that T-Apriori algorithm can successfully extract temporal association rules which described the close relationship between environmental factors and noctiluca scintillans events.

## 1. INTRODUCTION

In many practical areas, such as cadastre, meteorology, telecom, finance, plenty of time information exists in their data, that is to say their data has temporal relativity. The database storing temporal information can be named as temporal database. The former association rule mining ignores the time characters of things, however, the application areas are always changing with time(Ren Jiadong *et al*, 2004), for example, the transaction time in supermarket sale records, the brain wave of one patient, the number of browsing Web pages everyday, etc(Zhang Baowen *et al*, 2002). Temporal data reflects the development process of things which is of great benefit to uncovering the essence of things evolution. Temporal association rule mining is to discover the valuable relationship among the items in the temporal database.

Till now, the research and practise of temporal association rule mining have been popular, and many literatures have reported (H Mannila *et al,* 1997; Jef Wijsen *et al*, 1997; Sridhar Ramaswamy *et al,* 1998; Anthony K. H. Tung *et al*, 1999; Juan M.Ale *et al*, 2000; Sherri K. Harms *et al,* 2002; Sherri K. Harms *et al,* 2004). Here, we propose a new algorithm: T-Apriori based on time constraint and try to look for a new application of temporal association rule mining from a different viewpoint. Finally, we believe temporal association rule mining should be beneficial to the analysis of ecological phenomena generating and vanishing process. Since ecological phenomena have strong spatial and temporal characteristics, they usually occur in the same place while at different time and have different forms. For example, red tide phenomena of noctiluca scintillans often happens at the specific sea areas, but at different periods (noctiluca scintillans usually reproduce from march to may in a year, which makes the density of noctiluca scintillans increase sharply and results in red tide). We develop the concept of sequence of ecological events and apply T-Apriori algorithm to the analysis of ecological phenomena generating and vanishing process.

The remainder of the paper is organized as follows. Chapter 2 examines the concepts of association rule mining and temporal association rule mining. Chapter 3 discusses the temporal association rule mining methodology and T-Apriori algorithm. The meanings of ecological event and sequence of ecological event are explained in Chapter 4. Then a case is given in Chapter 5. Finally, Chapter 6 concludes the paper.

## 2. RELATED WORK

Association rule mining raised by Rakesh Agrawal is an important research problem in data mining field. Association rule mining aims at detecting the relationship of tuples in transactional database and serving decision making (Rakesh Agrawal et al, 1993). The association rule is the pattern knowledge existing in the given dataset or database. Let's data itemset $X = \{X_1, X_2,..., X_m\}$, $Y = \{Y_1, Y_2, ..., Y_n\}$, association rule can be represented as the form of "$X \rightarrow Y$", like "$X_1 \cap X_2 \cap ... \cap X_m \rightarrow Y_1 \cap Y_2 \cap ... \cap Y_n$"; $X_i$ ($i \in \{1…,m\}$) and $Y_j$ ($j \in \{1…,n\}$) are all in attribute-value format, and $X \cap Y = \Phi$. Association rule $X \rightarrow Y$ expresses "tuples satisfying conditions in X also satisfy conditions in Y "(Zhu Ming, 2002). If rule $X \rightarrow Y$ is true, then it has the support and confidence values, which can be described as:

* Corresponding author: ZHAI Liang, PhD candidate. His research interests include fundamental GIS, spatio-temporal database and data mining etc.

$$\text{Support } (X \rightarrow Y) = P\ (X \cup Y) \quad (1)$$

$$\text{Confidence } (X \rightarrow Y) = P\ (Y|X) \quad (2)$$

In essence, association rule mining is to find the rule sets satisfying minimum support threshold and minimum confidence threshold in the dataset or database (Rakesh Agrawal et al, 1993; Zhu Ming, 2002).

Comparing with the former association rule mining, temporal association rule adds time constraint (it can be time point or time range) on association rule. A transaction with time information can be described as: {TID, $I_1$, $I_2$ …$I_n$, $T_s$, $T_e$}. TID is the ID for each transaction; n-itemsets means there are n items in the itemset; $T_s$ and $T_e$ represent the start and the end of valid time respectively (or the start and the end of transaction). Valid time means the event occurring time, while transaction time the database time. $T_s$ may equal $T_e$, such as sale records in the supermarket (the transaction occurs at one moment). According to the definition of strong association rule "association rule strictly satisfies minimum support threshold and minimum confidence threshold", we can give the definition of strong temporal association rule.

**Definition 1** Let min_s and min_c represent minimum support threshold and minimum confidence threshold respectively, if and only if during [$t_s$, te], support ≥ min_s, confidence ≥ min_c, rule X→Y is a temporal association rule, which could be described as X→Y (support, confidence, [$t_s$, $t_e$]).

Itemset means the collection of items. If there were k items, we call it k-itemsets. The itemsets that satisfies min_s is called frequent itemset (Zhu Ming, 2002).

## 3. TEMPORAL ASSOCIATION RULE MINING

### 3.1 Methodology

No matter what kinds of tools or algorithms you select, strong temporal association rule mining can be divided into 4 steps. Step 2 and 3 are the core of the whole process, see Figure 1.

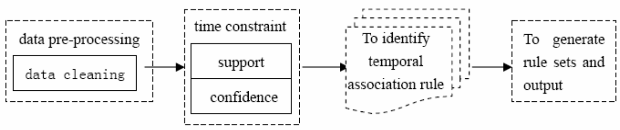


Figure1. Steps of Temporal Association Rules Mining

(1) Data pre-processing. It includes data cleaning, data integration, data exchange and data reduction. In virtue of data pre-processing, we can get high quality data mining object. This is an important step of data mining;

(2) To find the frequent itemsets which have the support no less that min_s;

(3) To generate association rules with frequent itemsets. It is different to generate association rules without time, because it adds time information on frequent itemsets. So here the association rules are temporal ones;

(4) To generate rule sets and output.

### 3.2 T-Apriori Algorithm Analysis

After introducing of the concepts and methodology of temporal association rule mining, we will focus on the algorithm for temporal association rule mining, Appriori algorithm is a basic one for getting frequent itemsets as well as an influential one in association rule mining. Through generating frequent itemsets with time information like temporal constraint, cycle and trend characteristics, we propose a temporal association rule mining algorithm: T-Apriori as to temporal dataset or database.

According to **Definition 1**, temporal association rule can be described as "X→Y(support, confidence, [$t_s$,$t_e$])". T-Apriori algorithm refers time as a constraint. First of all, we need analyze the temporal database with respect to time threshold. Time threshold is the time point or time range. Time range can be expressed as [min_$t_s$, min_$t_e$], while time point [min_t], for example, [1996, 2000] represents from 1996 to 2000; [2001.03] represents Mar.,2001. Then we should delete the time information in the temporal dataset or database in order to decrease computational complexity and apply the T-Appriori algorithm to generate frequent itemsets and corresponding temporal association rules. The followings are the details of T-Appriori algorithm.

#### 3.2.1 Generation of Frequent Itemsets

In T-Appriori algorithm, the process of frequent itemsets generation is similar to Appriori algorithm, but we need special treatment of time information. Here T is temporal database, $L_k$ is the frequent itemset.

**Input:** T, *min_s*

**Output:** *Results*=$\bigcup_k L_k$

**Algorithm Process:**

```
forall RecordSets do
  // RecordSets ∈T, a record set with time information
        Subtract(RecordSets)
  // to get TRecordSets satisfying time threshold
end
ItemSets= TRecordSets without TRecordSets.time
  // for easy calculation to delete time information
C1 = { Candidate 1-ItemSets }, L1 = { c ∈C1 | c.count ≥ min_s}
for ( k=2; Lk-1 ≠ φ; k++ )     do begin
  // untill no more frequent itemsets generate
      Ck = Apriori_Gen ( Lk-1 )
  // to generate k-item candiadte frequent itemsets
      forall transaction t ∈ItemSets    do begin
            Ct = subset ( Ck , t )
            forall Candidates c ∈Ct do
                  c.count++
  // to get the support for each candiadte frequent itemset
      end
      Lk={c ∈Ck|c.count ≥ min_s}
end
Results=∪k Lk
```

In this algorithm, c.count represent the support of frequent itemset c. *Apriori_Gen* function generates frequent itemset:$C_k$. The input parameter of *Apriori_Gen* is $L_{k-1}$, (k-1)-itemset and the output is $C_k$.In fact, $C_k$ is a superset of $L_k$. The generation of $C_k$ can be divided into two steps: Join and Prune; After the generation of $C_k$, we need scan the database and calculate the support of each subset of $C_k$. *subset* performs this function. For further information, see reference 5 (Rakesh Agrawal *et al*, 1994).

**Subtract function:** the input parameters are *RecordSets* and *time threshold*, it returns *TRecordsets* satisfying time threshold.
**Input:** *RecordSets*, $[min_t_s, min_t_e]$或$[min_t]$
**Output:** *TRecordSets*
**Algorithm Process:**

(1) **forall** *RecordSets* **do**
(2) **if** *RecordSets.time* $\in [min\_t_s, min\_t_e]$ or *RecordSets.time* $\in [min\_t]$ **then**
(3) *TRecordSets=RecordSets*
(4) **end**

### 3.2.2 Generation of Temporal Association Rule

**Input:** $\bigcup_k L_k$, $min\_c$
**Output:** rules like $x \rightarrow y(support, confidence, [t_s, t_e])$ or $x \rightarrow y(support, confidence, min\_t)$
**Algorithm Process:**

(1) **forall** $L_k$, $k \geqslant 2$ **do begin**
(2) $H_1$ = {*consequents of rules derived from* $L_k$ *with one item in the consequent*};
(3) **call** *ap-genrules* $(L_k, H_1)$;
(4) **call** *Add*$(L_k)$;
(5) **end**

**ap-genrules:** to generate rules, see reference 5 (Rakesh Agrawal *et al*, 1994) for detail information.

**Add:** the input parameter is frequent itemset $L_k$ (without time information), it needs scan the *TrecordSets* and returns the temporal association rule.

(1) **forall** *TRecordSets(i)*, $i \geqslant 0$ **do**
(2) **forall** $L'_k(k)$, $k \geqslant 2$ **do**
(3) **if** $L'_k \in TRecordSets(i)$ **then**
(4) *TRecordSets (i).time*$\cap(L_k - h_{m+1}) \rightarrow h_{m+1}$ *with confidence=conf and support=support*$(L_k)$
(5) **end**
(6) **end**

## 4. ECOLOGICAL EVENT

The regional spatio-temporal distribution of populations of organisms is intimately controlled by local environmental characteristics, including physical, chemical and meteorological aspects. At different time, the change of environmental characteristics will lead to some ecological phenomena happening at same place or different places. Ecological event refers to the behavior of living organisms under certain spatio-temporal states. All environmental factors, time and the behavior of living organisms consist the attribute set of an ecological event. A sequence of ecological events $\boldsymbol{s}$ can be expressed as a trituple ($\boldsymbol{s}$, $t_s$, $t_e$), $\boldsymbol{s} = \{(A_1, t_1), (A_2, t_2) \dots (A_n, t_n)\}$, is an ordered ecological events set; $A_i$ (i = 1, 2, … n) represents ecological event, its attributes $<attribute_{i1}, attribute_{i2}, \dots attribute_{im}>$; $t_s$, $t_e$ represents starting and ending time; $t_i$ is the occurring time of an event, $t_s \leqslant t_i < t_e$, $t_i < t_{i+1}$ (i = 1, 2, … n-1). For example, Figure2 shows a sequence of ecological events $\boldsymbol{s}$ = ($s$,*1998.01, 1998.09*), s = {($A_1$, *1998.01*), ($A_2$, *1998.02*), ($A_3$, *1998.03*), ($A_4$, *1998.05*), ($A_5$, *1998.08*), ($A_6$, *1998.09*).

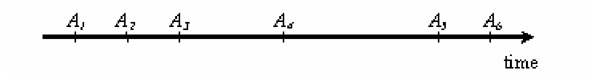


Figure2. A Sequence of Ecological Events

The red tide phenomena is a kind of typical ecological event, for example, in Mar. 1992 in the Dapeng bay, South China Sea, sea water temperature (Tem) is in the range of 19-22℃, salinity *(Sal)* 30-35, red tide A occurs, which could be described as (red tide A, Mar.1992), attributes of A$<Tem, Sal>$, $Tem \in [19, 22]$, $Sal \in [30, 35]$.

In order to find the main environmental factors of ecological event and the association relationship among these environmental factors, we need form a decision table (in the format of transactional records) (Fenzhen Su *et al*, 2004). According to prior knowledge, we assume that the behavior of living organisms is affected by factors $f_1, f_2, \dots f_n$, at different time $t_1, t_2, \dots t_n$, $f_1, f_2, \dots f_n$ has different values, Y or N indicates the binary presence or absence state of living organisms. The long term behavior of these living organisms can be regarded as a sequence of ecological events and form a decision table including values of the attribute and states, see Table1.

| ID | time | $f_1$ | $f_2$ | … | $f_n$ | Y/N |
|---|---|---|---|---|---|---|
| 1 | $t_1$ | $f_{11}$ | $f_{21}$ | … | $f_{n1}$ | Y |
| 2 | $t_2$ | $f_{12}$ | $f_{22}$ | … | $f_{n2}$ | N |
| … | … | … | … | … | … | … |

Table1. Decision Table

Prior research indicates that the phenomena of red tide and its life process are not some isolate events, but rather an interactional process. In recent 15 years, the frequency and scale of red tide are increasing continuously, which seriously destroyed the ocean ecological system and brought damages to fishing resources and mariculture. Furthermore, some types of hazardous red tide may threaten people live through food chain (Shen Fei *et al*, 2003). Scientific researchers have long observed and recorded physical, chemical and meteorological parameters in corresponding ocean areas which contains abundant ecological information. We should adopt data mining techniques via statistical analysis in detecting main environmental factors of red tide and building association relationship among these factors, which would help us in understanding the generation mechanism of red tide and provide references for red tide monitoring and management. In addition, red tide phenomena have strong temporal characteristics and vary with seasons (Qi Yuzao, 2003). So temporal association rule mining seems to be an ideal selection. We call problems like this as temporal association rule mining based a sequence of ecological events. The following will elaborate on this with an example of red tide events in Dapeng bay, South China Sea.

## 5. CASE STUDY

### 5.1 Data Resource and Data Pre-processing

#### 5.1.1 Data Resource

Data resource is the red tide monitoring data in Dapeng bay. Dapeng bay lies between Dapeng byland (Baoan county, Guangdong province) and Jiulong byland (Hongkong) in the range of latitude (norht) 22°24′-22°36′, longitude (east) 114°12′-114°30′, see Figure3. In Figure3, S0, S1 and S2 represent the monitoring stations (Huang Changjiang *et al*, 1997). Dapeng bay is one of the sea areas often subject to red tide in the South China Sea and Noctiluca scintillans is a type of popular red tide organisms in Dapeng bay. Noctiluca scintillans is harmful to seawater ecological balance as well as to fishing resources and mariculture (Huang Weijian *et al*, 1993). Table 2 consisting of observing data from (Qi Yuzao, 2003; Huang Changjiang *et al*, 1997) is similar to a decision table.

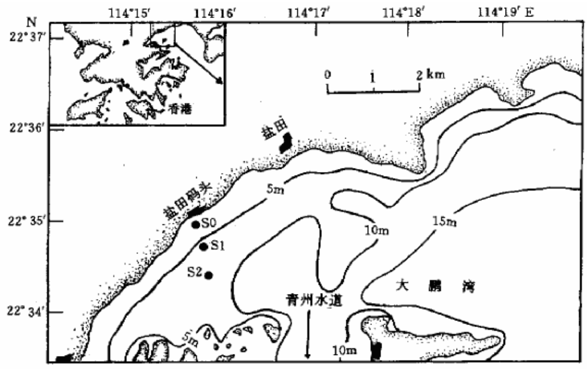


Figure3. Red Tide Monitoring Stations in Dapeng Bay

| ID | time | Principal Environmental Parameters Measurements | | | | | | phenomena of noctiluca scintillans |
|---|---|---|---|---|---|---|---|---|
| | | seawater temperature /℃ | salinity | dissolved oxygen /(mg·L$^{-1}$) | pH | chemical oxygen consumption /(mg·L$^{-1}$) | chlorophyll a /(μg·L$^{-1}$) | |
| 1 | 1991.01 | 19.0 | 32.800 | 7.27 | 8.00 | 0.30 | 1.4 | no |
| 2 | 1991.02 | 17.4 | 32.466 | 7.28 | 8.27 | 0.33 | 0.6 | no |
| 3 | 1991.03 | 19.8 | 31.974 | 7.55 | 8.19 | 0.50 | 2.1 | yes |
| 4 | 1991.04 | 20.1 | 32.452 | 6.84 | 8.15 | 0.73 | 1.0 | yes |
| 5 | 1991.05 | 25.6 | 32.861 | 6.80 | 8.18 | 0.48 | 0.7 | no |
| 6 | 1991.06 | 28.6 | 32.452 | 6.22 | 8.17 | 0.43 | 0.8 | no |
| … | … | … | … | … | … | … | … | … |

Table2. Environmental Parameters Measurements and Red Tide Phenomena of Noctiluca Scintillans during 1991 and 1992 in Dapeng Bay (part)

### 5.1.2 Data Pre-processing

Since the data in table 2 is quantitative, it is difficult to use association rule mining techniques directly (Ramakrishnan Srikant *et al*, 1996; Ma Chaofei *et al*, 2003; Jeremy Mennis *et al*, 2005). Ramakrishnan Srikant *et al* define the problem of mining association rules over quantitative and categorical attributes in large relational tables as the Quantitative Association Rules problem (Ramakrishnan Srikant *et al*, 1996). In order to increase the data quality and mining efficiency, we should partition the values of the attribute so as to mapping the quantitative association rule problem into the Boolean association rule problem. Ramakrishnan Srikant *et al* proposed an equi-depth partitioning the values of the attribute method and a formula to compute partitioning number. Equi-depth partitioning may not work very well on highly skewed data and is inclined to generate too many similar rules (Ramakrishnan Srikant *et al*, 1996). The data in Table2 varies much like seawater temperature and chlorophyll a, so it is not suitable to adopt this equi-depth partitioning method.

Clustering analysis is the searching for groups (clusters) in the data, in such a way that samples belonging to the same cluster resemble each other, whereas samples in different clusters are dissimilar. K-means algorithm is one of the most widely used clustering (partitioning) algorithms. This algorithm initially takes k clusters. Next, it examines each component in the population and assigns it to one of the clusters depending on the minimum distance. The centroid's position is recalculated every time a component is added to the cluster and this continues until all the components are grouped into the final k clusters (Zhai Liang *et al*, 2005). We can learn from the idea of clustering analysis and adopt k-means algorithm to partition the values of the attribute. According to the observation values of Dapeng bay and their distribution status, we can assign the partitioning number. We also need to point out the partitioning number should not be so big that there were few samples (observation values) in one cluster. Spss software is employed here do this job and we get Figure4, based on which we can map the quantitative association rule problem into the Boolean association rule problem and get Table3, a table in the format of transactional records. In Table3, $I_{uv}$ represents attribute items (u corresponds to “seawater temperature”, “salinity”, “dissolved oxygen”, “pH”, “chemical oxygen consumption” and “chlorophyll a” respectively, u∈[1, 6]; v corresponds to “class 1”, “class 2” and “class 3” respectively, v∈[1, 4]); “Y” and “N” represents whether the red tide event has occurred.

| | class 1 | class 2 | class 3 |
|---|---|---|---|
| seawater temperature | [16.6, 20.6] | [22.5, 26.6] | [27.2, 30.2] |

(a)

| | class 1 | class 2 | class 3 |
|---|---|---|---|
| salinity | [26.077, 27.248] | [29.402, 31.032] | [31.308, 33.108] |

(b)

| | class 1 | class 2 | class 3 |
|---|---|---|---|
| dissolved oxygen | [4.14, 6.33] | [6.80, 7.84] | [8.15, 9.21] |

(c)

| | class 1 | class 2 |
|---|---|---|
| pH | [8.0, 8.15] | [8.17, 8.31] |

(d)

| | class 1 | class 2 | class 3 | class 4 |
|---|---|---|---|---|
| chemical oxygen consumption | [0.18, 0.36] | [0.41, 0.52 ] | [0.54, 0.64] | [0.73, 0.88] |

(e)

| | class 1 | class 2 | class 3 |
|---|---|---|---|
| chlorophyll a | [0.6, 2.1] | [2.4, 3.5] | [4.4, 6.4] |

(f)

Figure4. Attributes Partition of Environmental Parameters Measurements

| ID | time | Item | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | seawater temperature | salinity | dissolved oxygen | pH | chemical oxygen consumption | chlorophyll a | Y/N |
| 1 | 1991.01 | $I_{11}$ | $I_{23}$ | $I_{32}$ | $I_{41}$ | $I_{51}$ | $I_{61}$ | N |
| 2 | 1991.02 | $I_{11}$ | $I_{23}$ | $I_{32}$ | $I_{42}$ | $I_{51}$ | $I_{61}$ | N |
| 3 | 1991.03 | $I_{11}$ | $I_{23}$ | $I_{32}$ | $I_{42}$ | $I_{52}$ | $I_{61}$ | Y |
| 4 | 1991.04 | $I_{11}$ | $I_{23}$ | $I_{32}$ | $I_{41}$ | $I_{54}$ | $I_{61}$ | Y |
| 5 | 1991.05 | $I_{12}$ | $I_{23}$ | $I_{32}$ | $I_{42}$ | $I_{52}$ | $I_{61}$ | N |
| 6 | 1991.06 | $I_{13}$ | $I_{23}$ | $I_{31}$ | $I_{42}$ | $I_{52}$ | $I_{61}$ | N |
| … | … | … | … | … | … | … | … | … |

Table3. Mapping to Transactional Records (part)

### 5.2 Applying T-Apriori Algorithm

Our T-Apriori algorithm is implemented in VC++6.0 and our were performed on an Dell PC with 256 MB of main memory and P Ⅳ 2.40GHz. The data from Table3 was stored in Microsoft Access database. Figure5 shows the preferences dialog and output results, for example let minimum support = 0.20; minimum confidence = 0.65; time from 1991.02 to 1992.05.

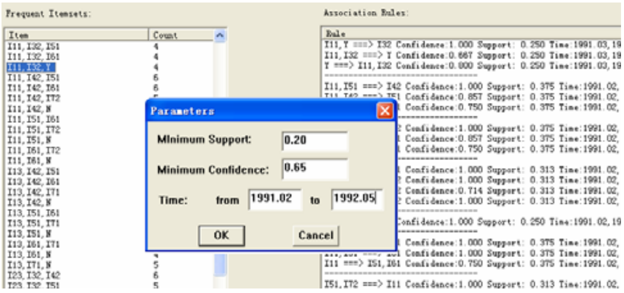


Figure5. Preferences Dialog and Output Results

Many rules could be extracted. It's difficult to identify the ones useful or valuable and we need further analysis or interpretation in virtue of our knowledge of red tide. For example, we can choose one rule with "minimum support = 0.20; minimum confidence = 0.65; time from 1991.02 to 1992.05" to analyse:

$$I_{11},I_{32}\rightarrow Y(0.25,0.667, [1991.03,1991.04]\cup[1992.03,1992.04])$$

This rule means that from the historical observation data (from 1991 to 1992) we could find out that there would occur red tide phenomena of noctiluca scintillans in the target sea area (Da peng bay) if seawater temperature lied in the range of 16.6-20.6 ℃ and dissolved oxygen in the range of 6.80-7.84 mg•L-1. It indicates that seawater temperature and dissolved oxygen contribute more to noctiluca scintillans events than other environmental factors, which could enlighten ecological scientists to research on the mechanisms of these two factors' influence. It also shows that red tide phenomena of noctiluca scintillans in Da peng bay always occur during March and April every year, which would remind relating departments to take special actions in these months.

Being affected by the data source, we only discuss the relationship between noctiluca scintillans phenomena and environmental factors in a given area and time period. The identification of noctiluca scintillans alarm value is changed with different time and sea areas. Here we just select a Boolean value (Y/N) to compute. If the data content were abundant, we would partition the noctiluca scintillans phenomena into different levels according to its density index, like serious, less serious and trivial. Thus, we could compute and analyze the influence of environmental factors variation on noctiluca scintillans density, from which we would understand the growing characteristics of noctiluca scintillans and take corresponding measures to forecast. With the accumulation of observation data year by year, we shall learn from the tremendous history data and get useful knowledge through spatio-temporal data mining techniques. It is critical to build the knowledge database of various algae in the establishment red tide forecasting and warning system.

## 6. CONCLUSIONS

With the development of technology, the difficulty to obtain data is decreasing and the difficulty to analyse the huge volume of data is increasing. That is to say, it is technically more difficult to extract knowledge from the huge volume of data which are currently being stored in databases (Fenzhen Su *et al*, 2004). Temporal association rule mining aiming at detecting correlations in transactional and relational data that possess a time component seems to be a promising analytical tool for spatio-temporal data analysis. We have presented in this paper a temporal association rule mining algorithm: T-Apriori algorithm and discussed the problems of temporal association rule mining based on a sequence of ecological events. Red tide phenomena of noctiluca scintillans during 1991 and 1992 in Dapeng bay, the South China Sea was taken as an example to validate T-Apriori algorithm and K-means clustering analysis was used to map the quantitative association rule problem into the Boolean association rule problem. Our example experiment shows that T-Apriori algorithm can successfully extract temporal association rules which describe the close relationship between environmental factors and noctiluca scintillans events. Although this case study shows that temporal association rule mining based on T-Apriori algorithm is effective in detecting the correlation between environmental factors and ecological events, there are still several issues that warrant further research.

First, the efficiency of T-Apriori algorithm should be taken into account for dealing with very large database. We shall test its efficiency and make improvements. Second, the impact of using clustering method in partitioning the values of the attribute so as to mapping the quantitative association rule problem into the Boolean association rule problem needs careful consideration, since this method may not be significant in other situations and the partition results will affect the generated rules directly. Third, how to extract automatically interesting rules from the multitude of rules is one puzzling issue. Even many business data mining softwares failed in solving this problem well. A number of authors have suggested approaches beyond support and confidence for measuring the "interestingness" of association rules (Jeremy Mennis *et al*, 2005). Finally, and perhaps most important, temporal association rule mining should be extended to sequential patterns mining, periodicity mining, cyclic rules mining, ratio rules mining, etc, for better understanding the ever changing world.

## ACKNOWLEDGEMENTS

The authors thank Mr. LV Huaxin (School of Resource and Environment Science, Wuhan University, China) for his assistance during the research. This research was partially funded by the open research subject of Key Laboratory of Geo-informatics of State Bureau of Surveying and Mapping and National Surveying and Mapping Fund (No. 1460130524207).

## REFERENCES

Anthony K. H. Tung, Hongjun Lu, Jiawei Han, Ling Feng. Breaking the Barrier of Transactions: Mining Inter-transaction Association Rules, 1999. In: International Conference on Knowledge Discovery and Data Mining, Diego, America, pp.297~301

Fenzhen Su, Chenghu Zhou, Vincent Lyne, Yunyan Du, Wenzhong Shi, 2004. A data-mining approach to determine the spatio-temporal relationship between environmental factors and fish distribution. Ecological Modelling, 174(4), pp. 421-431

Heikki Mannila, Hannu Toivonen, A. Inkeri Verkamo, 1997. Discovery of Frequent Episodes in Event Sequences. Data Mining and Knowledge Discovery, 1(3), pp.259~289

Huang Changjiang, Qi Yuzao, Huang Yihua, Lin Xiaotao, 1997. The Population Ecology and Causative Mechanisms of Red Tide of Noctiluca Scintillans in Dapeng Bay, the South China Sea. Oceanology and limnology, 28(3), pp. 245-255

Huang Weijian, Qi Yuzao, 1993, Environmental Influence on Growth of Noctiluca Scintillans in Dapeng Bay, North of South China Sea. Marine Science Bulletin, 12(2), pp. 17-22

Jef Wijsen, Robert Meersman, 1997. On the Complxity of Mining Temporal Trends. In: ACM SIGMOD Int. Workshop on Research Issues on Data Mining and Knowledge Discovery, British Columbia, Canada, pp.77~84

Jeremy Mennis, Jun Wei Liu, 2005. Mining Association Rules in Spatio-Temporal Data. Transactions in GIS, 9(1): pp. 5-17

Juan M.Ale, Gustavo H. Rossi, 2000. An Approach to Discovering Temporal Association Rules. In: 2000 ACM Symposium on Applied Computing (SAC'00). Como, Italy, pp.294~300

Ma Chaofei, Liu J ianqiang, 2003. Multidimensional and Quantitative Association Rule Mining from Remote Sensing Images. REMOTE SENSING TECHNOLOGY AND APPLICATION, 18(4), pp. 243-247

Qi Yuzao, 2003. China Ocean Red Tide. Science Press, Beijing, pp: 189-221

Rakesh Agrawal, Tomasz Imielinski, Arun Swami, 1993. Mining Association Rules between Sets of Items in Large Databases. In: ACM SIGMOD Int. Conference on Management of Data. Washington DC, USA, pp. 207-216

Rakesh Agrawal, Ramakrishnan Srikant, 1994. Fast Algorithms for Mining Association Rules. In: the 20th International Conference on Very Large Databases (VLDB). Santiago, Chile, pp. 487-499

Ramakrishnan Srikant, Rakesh Agrawal, 1996. Mining Quantitative Association Rules in Large Relational Tables. In: ACM SIGMOD Int. Conference on Management of Data. Montreal, Canada, pp. 1-12

Ren Jiadong, Ren Dongying, 2004. Mining Multi-level Temporal Association Rules from Distributed Timestamp Databases. Computer Engineering, 30(16), pp. 63-64

Shen Fei, Wang Hongzi, Feng Jianfeng, Sun Jing, 2003. Application of Knowledge Discovery in the Study on the System of Forecasting and Warning of Red Tide. Ocean Technology, 22(2), pp.19-22

Sherri K. Harms, Jitender Deogun, Tsegaye Tadesse, 2002. Discovering Sequential Association Rules with Constraints and Time Lags in Multiple Sequences. In: The 13th International Symposium on Methodologies for Intelligent Systems. Lyon, France, pp.432~441

Sherri K. Harms, Jitender S. Deogun, 2004. Sequential Association Rule Mining with Time Lags. Journal of Intelligent Information Systems, 22(1) , pp. 7~22

Sridhar Ramaswamy, Sameer Mahajan, Avi Silberschatz,1998. On the Discovery of Interesting Patterns in Association Rules. In: International Very Large Databases Conference, New York, America, pp.368~379

Zhai Liang, Tang Xinming, Wang Zhongyuan, Wu Lan, 2005. Analysis of 1:50 000 Database Content Survey Results. Geomatics and Information Science of Wuhan University, 30(4), pp. 362-365

Zhang Baowen, He Huachan, 2002. Progress of Temporal Data Mining Research. Computer Science, 29(2), pp.124-126

Zhu Ming, 2002. Data Mining. University of Science and Technology of China Press, Hefei, pp. 11-12