Mining time-series association rules from Western Pacific spatial-temporal data

Weixuan Ma\textsuperscript{1,2}, Cunjin Xue\textsuperscript{1,3}, Junqi Zhou\textsuperscript{2}

\textsuperscript{1}Institute of Remote Sensing and Digital Earth Chinese Academy of Sciences, No.9 Dengzhuang South Road, Haidian District, Beijing, 100094, China

\textsuperscript{2} Wuhan University, Luojia Hill, Wuchang District, Wuhan 430072, Hubei, China

E-mail: cjxue@ceode.ac.cn

Abstract. With increasing concerns about the environmental problem as well as tremendous environmental issues impacting on our daily life, a new requirement for analysis of environmental changes and effect has been proposed. In this paper we use Western Pacific events and basic background database as its data source to find the association between different marine parameters. The improved Apriori algorithm is utilized to discover knowledge in magnanimous spatio-temporal data. There are two main steps. First is according to the different variation degree of each point, the study area can be divided into lots of spatial-temporal transaction zones. Second is use the improved Apriori algorithm for spatial-temporal data mining. For the need of mining algorithm, the quantitative attributes need to be transformed into qualitative attributes. The concept generalization method is utilized to divide the original attribute data into several levels. Then the Apriori algorithm can be used to discover the potential association between marine parameters within the given time frame.

1. Introduction

The El Nin\textdegree o Southern Oscillation (ENSO) phenomenon is a result of complex interaction between the atmosphere and the hydrosphere in the tropical Pacific \cite{1}. It is one of the most significant sources of interannual climate variability on Earth \cite{2}. Many observations and studies suggested that the marine attributes such as precipitation, sea surface temperature (SST), wind, sea level anomalies (SLA), chlorophyll (CHL), are the parameters closely associated with ENSO. The goal of this paper is to find the relationship between ocean attributes. In addition to discovering their association reflect by their spatial and temporal variation. For the past years, many countries have developed different ocean-observatory systems for continuously monitoring and collecting the oceanic data. Large amounts of spatial data with temporal information was collected through remote sensing. These data provide a way for scientists to study the trends and variability of marine attributes for different regions. But by study the change of observed data can only reflect the variety of single attribute, the association between different attributes is implied. This situation makes it crucial to develop tools for the discovery of interesting knowledge between different objects from large spatial-temporal databases. Many approaches have been proposed to extract information, and the data mining is one of the most important methods \cite{3}. Data mining extracts implicit, previously unknown, and potentially useful information from databases. These potential and deeply hidden information and knowledge may

\textsuperscript{3} To whom any correspondence should be addressed.
help people have a deeper understanding of inherent nature in geographic phenomenon and process, also the generation, development and variation of these phenomena. This approach was introduced in the mid 1990s to discover patterns that occur frequently in a sequence database [4]. A typical use of data mining is market-basket analysis.

The majority of the data mining algorithms was developed for the analysis of relational and transactional databases. A famous classic algorithm named Apriori is often used to discover frequent itemsets in transactions. When it is used in the Western Pacific spatial-temporal data mining there are still several key issues need to solve. The first is to transform the spatial-temporal data into transactions [5]. The traditional applications, such as market analysis, decision support, fraud detection, and business management, transactions are independent from each other. For the spatial data, there is no explicit boundary to transform into transactions. Furthermore, the spatial-temporal relationship is exist between transactions in marine data. The second is the classical Apriori algorithm is not suitable for finding spatial-temporal patterns. In the traditional market-basket analysis, people just want to get the information like "if bread's sales volume rise, then milk's sales volume rise ". For our issue we prefer to get the knowledge as "if CHL heavily drop in zone one, then precipitation heavily rise in zone two after three months ". These two rules are very different. Compared with the former rule, the two events in the latter rule belong to different transactions and temporal and spatial information is include. Third, the goal of this paper is to discover associations between abnormal variation between ocean attributes in different area. Rectangle tool is employed to mark several centers of value divergence. These rectangles divide the whole Western Pacific area into several small pieces. The ocean data in each pieces can transform into transactions while using a region ID to maintain spatial information. A series of transactions composed a data sequence. A set of data sequences formed the sequence database which is ordered by time sequence. Then an improved Apriori algorithm for mining inter-transaction association rules [6] that include context information is proposed. Simultaneously, in the process of the algorithm, combined with the characteristic of abnormal attribute data we apply a Gaussian distribution support degree to reduce the number of candidates.

The remainder of this paper is organized as follows. Section 2 presents the data preprocessing process. Section 3 introduce the improved Apriori algorithm. Some experimental results are given in section 4. Section 5 draws a short conclusion.

2. Data preprocessing

2.1. Data source
The study area we focus on in this paper is the Western Pacific which locate in the longitude 120 degrees east to 80 degrees west and latitude 40 degrees north to 40 degrees south. The initial remote sensing data we used in this paper is as follows: the global ocean surface temperature data is monitored by NOAA satellite [7]. The SeaWiFS sensor continues to observe the ocean in the visible and near infrared spectrum, and it provides valuable data on the marine biosphere. We use the concentration of chlorophyll-a here. From satellite TRMM we get the precipitation data. The temporal resolution of these data is monthly.

2.2. Transaction definition
After conduct empirical orthogonal function (EOF) and Morlet wavelet analysis process on original data, it will be easy for us to discover several centers of value divergence in each marine attribute on the map. Then we use the rectangle tool to choose some interesting areas. Since our goal is to study the association between abnormal events, select the centers of divergence will be more representative. Figure 1 shows the result of sea surface chlorophyll concentration object after EOF processing. In this figure, we pick up six curious zones from the whole Western Pacific area and mark each of them with Zone1 to Zone 6. Table 1 shows the spatial information transformed from Figure 1. The accurate position information is saved in the spatial information table and connect with other tables in the database through Object ID.
After dividing the entire area into several small zones, each rectangle can be treated as a transaction. The raw data after departure processing can reflect the abnormal variation more clearly. Table 2 shows part of the abnormal variation data transformed from initial data.

Table 1. The spatial information from Figure 1

| Zone   | Object ID | Spacial scale                        |
|--------|-----------|--------------------------------------|
| Zone1  | 10001     | 26° N-35° N, 159° W-180° W           |
| Zone2  | 10002     | 1° N-15° N, 132° E-155° E            |
| Zone3  | 10003     | 5° S-10° S, 160° E-175° E            |
| Zone4  | 10004     | 30° S-35° S, 158° W-175° W           |
| Zone5  | 10005     | 18° S-2° N, 119° W-172° W            |
| Zone6  | 10006     | 4° N-13° N, 90° W-165° W             |

Table 2. Abnormal variation of ocean attributes in the database

| Date   | Abnormal CHL variation in 10001 | Abnormal SSH variation in 10001 | Abnormal CHL variation in 10002 |
|--------|---------------------------------|---------------------------------|---------------------------------|
| 2003/1 | 0.27                            | -0.29                           | 0.34                            |
| 2003/2 | 0.44                            | 0.34                            | 0.31                            |
| 2003/3 | 0.13                            | 0.11                            | 0.28                            |
| ...    | ...                             | ...                             | ...                             |

2.3. Quantitative attributes discretization
The transaction database contains a series of attributes. Traditional mining algorithms only concern the co-occurrence of different items in the same transaction and the quantitative information is ignored. To this problem Srikant and Agrawal [8] proposed an approach by partitioning the quantitative attribute domains, combining adjacent partitions, and then transforming the problem into binary problem. Afterwards, Georgii, Richter, Ruckert and Kramer [9] proposed a method without discretization. There is an other approach [10] using a standard statistical measure to consider the distribution of the continuous data directly. In this paper we employ a method to partition continuous data into domains.
Here is an example to convert the data in Table 2 into intervals. Data in the same column can figure out a variance value marked $\sigma$. According to the relationship between attribute value and variance $\sigma$ in Table 3 we can get the result shown in Table 4.

The quantitative attributes have been transformed into disjoint intervals and the mining algorithm can be used to discover association rules.

**Table 3. The mapping intervals for attribute quantitative value**

| Flag value | Meaning            | Mapping intervals |
|------------|--------------------|-------------------|
| 2          | Severe positive changes | value $\leq \sigma$ |
| 1          | Slight positive changes   | $0.5\sigma \leq \text{value} < \sigma$ |
| 0          | No changes            | $-0.5\sigma \leq \text{value} \leq 0.5\sigma$ |
| -1         | Slight negative changes | $-1.0\sigma \leq \text{value} \leq -0.5\sigma$ |
| -2         | Severe negative changes | value $\leq -\sigma$ |

**Table 4. The mapping result from Table 2**

| Date      | Abnormal CHL variation in 10001 | Abnormal SSH variation in 10001 | Abnormal CHL variation in 10002 | ... |
|-----------|----------------------------------|----------------------------------|----------------------------------|-----|
| 2003/1    | 2                                | -2                               | 2                                | ... |
| 2003/2    | 2                                | 2                                | 2                                | ... |
| 2003/3    | 2                                | 1                                | 2                                | ... |
| ...       | ...                              | ...                              | ...                              | ... |

3. Spatio-temporal association mining algorithm

An inter-transaction association rule is spanned across m intervals to find if an association exists between the transactions. The boundary between transactions is broken. In this section we modify the well-known Apriori algorithm to find the associations include contextual information among precipitation, CHL, SST, SLA and wind in the Western Pacific data.

The mining algorithm comprise two main procedures. One is the discovery of frequent itemsets. Another is generate candidate items. We improve the traditional Apriori algorithm in two main aspects: the first is modify the method to generate frequent items $L_k$. Second is a new support value strategy to evaluate frequent itemsets. The next two subsections is to introduce the algorithm around these two aspects.

3.1. Normal distribution support

Association rule mining uses the concepts of support and confidence. The support is the probability of a record in the database satisfying the set of predicates contained in both the antecedent and consequent [11]. The infrequent sequences will be removed in each $L_k$. It is very important in reducing the size of $C_k$ and improve the mining efficiency, but an improper support may generate a lot of useless items or destroy interested items.

The purpose of this paper is to mining the association between abnormal events. Serious anomalies is what people pay close attention to. These events have important research value but rarely happen, if using a same support for different objects, the extreme events may disappear. In this paper we use the normal distribution support, for normal event use of a higher support, extreme events with lower support. In this way the extreme items can be retained, and also the efficiency of the algorithm is improved.

3.2. Discover frequent itemsets
Definition 1. Let $I = \{i_1, i_2, \ldots, i_j\}$ be a set of items. Let $D$ be a dimensional attribute and $\text{Dom}(D)$ be the domain of $D$. A transaction database is a database containing records in the form $(d, I_j)$, where $d \in \text{Dom}(D)$ and $I_j \subseteq I$.

The basic principle of Apriori is: the first pass of the algorithm is simply counts the item occurrences to determine the large 1-itemsets. First, the large itemsets $L_{k-1}$ that found in the $(k-1)$th pass are used to generate the candidate itemsets $C_k$, by using the apriori-gen function. Next, scan the database and count the support of each candidates in $C_k$. The algorithm will be ended until the candidate large $k$ itemset $C_k$ is empty [12].

In the first phase, $L_1$ is found from $C_1$. $C_1$ can be generated by listing all the distinct items in databases. We can directly generate $L_1$ by scanning the database and computing supports of each $C_1$. Traditional $C_2$ can be obtained by directly joining $L_1$ with $L_1$, but here time-interval relations must be generated. For example, the time-interval is 2, $L_1 = \{\text{CHL}, \text{SSH}\}$, then $C_2 = \{\text{(CHL, SSH*-2)}, \text{(CHL, SSH*-1)}, \text{(CHL, SSH*0)}, \text{(CHL, SSH*1)}, \text{(CHL, SSH*2)}\}$. We use the symbol "*" with an number to express the time-interval between current item with the former one. When computing the supports for $C_2$, we should not only find items in the same transaction, but also the transactions before and after the current transaction inside time-interval. Figure 2 depicts the process to compute the support counts.

The next question is how to generate $C_k$. If $(I_1, I_2, I_3)$ is a frequent sequence in $L_3$, then the two sequences $(I_1, I_2)$ and $(I_2, I_3)$ must be also frequent. Therefore, if the sequence $(I_1, I_2)$ and $(I_2, I_3)$ exist in $L_2$, $(I_1, I_2, I_3)$ must exist in $C_3$. It can be generalized to generate $C_k$ from $L_{k-1}$.

4. Conclusion

This paper uses an improved Apriori algorithm to mining inter-transaction rules over the Western Pacific data to find the associations between marine attributes. In order to get the mining result with
contextual temporal and spatial information, first we use the EOF and Morlet wavelet analysis process to divided the whole area into small regions, and transform the data in each small regions into transactions. Second is converting the quantitative data into intervals, in this way the quantitative problem can be transformed into a binary problem and then the Apriori algorithm can be used. In the process of mining, a normal distribution support value is used to maintain the extreme event. The boundary between transactions is broken, the association rules with spatial and temporal information are found.

5. Acknowledgement
The research is supported by National Key Basic Research Program of China (project No. 2009CB723903), combined with National Natural Science Foundation of China (project No. 40901194), National High Technology Research and Development Program of China (project No. 2012AA12A403-5), and the director projects No.Y2ZZ06101B and No.Y2ZZ18101B supported by CEODE.

References

[1] Cane M A 2005 The evolution of El Nino past and future Earth and Planetary Science Letters 230 227
[2] Philander G S 1989 El Nino, La Nina, and the Southern Oscillation (San Diego: Academic Press)
[3] Frawley W J, Piatetsky Shapiro G and Matheus C J 1991 Knowledge Discovery in Database: An Overview (Cambridge, MA: AAAI/MIT Press)
[4] Agrawal R and Srikant R 1995 Mining sequential patterns Proc. Int. Conf. on Data Engineering(Taipei: IEEE Computer Society) pp 3–14
[5] Han J, Koperski K and Stefanovic N 1997 A system prototype for spatial data mining Proc. the ACM SIGMOD Int. Conf. on management of data (Tucson, AZ) pp 553–556
[6] Feng L, Dillon T and Liu J 2001 Inter-transactional association rules for multi-dimensional contexts for prediction and their application to studying meteorological data Data and Knowledge Engineering, 37 115
[7] Santos A M P 2000 Fisheries oceanography using satellite and airborne remote sensing methods: a review Fisheries Research 49 1
[8] Srikant R and Agrawal R 1996 Mining quantitative association rules in large relation tables Proc. the ACM SIGMOD Int. Conf. on management of data (Montreal, Canada)
[9] Georgii E, Richter L, Ruckert U and Kramer S 2005 Analyzing microarray data using quantitative association rules Bioinformatics 21 123
[10] Aumann Y and Lindell Y 2003 A statistical theory for quantitative association rules Journal of Intelligent Information Systems 20 255
[11] Koperski K and Han J 1995 Discovery of spatial association rules in geographic information databases Proc. the Fourth Int. Symposium on Large Spatial Databases pp 47–66
[12] Agrawal R and Srikant R 1994 Fast algorithms for mining association rules Proc. the 20th Int. Conf. on Very Large Databases (Santiago: IBM Almaden Research Center)