Design of ETL Process on Spatio-temporal Data and Study of Quality Control

Buyu Wang, Changyou Li, Xueliang Fu, Meian Li, Dongqing Wang, Huibin Du, Yajuan Xing

Design of ETL Process on Spatio-Temporal Data and Study of Quality Control[1]

Buyu Wang[1], Changyou Li[1], Xueliang Fu[1], Meian Li[1], Dongqing Wang[1], Huibin Du[1], Yajuan Xing[2]

[1]Inner Mongolia Agricultural University, Hohhot, 010018, P. R. China
[2]Inner Mongolia Fengzhou Vocational College, Hohhot, 010018, P. R. China
E-mail: bywang05@163.com

Abstract. In order to use the space-time data mining technology to conduct operation research in WuLiangSuHai Eutrophication, the water quality sensor parameters of heterogeneous data which reflect the characteristics should set up a spatial data warehouse through ETL process, and water quality sensors for quality control of spatial and temporal data plays a vital role in building an effective analytical environment. The paper designs the ETL process from the data and water quality sensors artificial duty and other heterogeneous data sources spatial data, and proposes data quality control strategy based on the incremental frequency rule engine and the space the inverse distance weighting on the Combination. Experiments show that the incremental frequency rule engine could more effectively find the missing sensor data and abnormal, Space inverse distance weighting method can find the missing data and outliers in the errors within the allowed interpolation processing, ETL procedure is effective and feasible.

Keywords: Sensor data, Frequency increment rule engine, Inverse distance weighted method

1 Introduction

At present, the integrated analysis and process of water quality services have been investigated using the spatial data mining techniques [1], [2], [3]. In particular, some work has studied multi-eutrophication services of water environments based on the heterogeneous data sources [4]. The pre-requisite for the investigation of eutrophication related services is to establish a data warehouse including the water-quality data characterized by a variety of eutrophication attributes. The design of the ETL (Extract, Transform, Load) procedure plays a key role to establish such an effective analysis environment.

ETL process design, many scholars from the conceptual model, conceptual model to logical model of the transformation of both done a lot of research work [5], [6]. Abnormal data, some scholars have proposed the use of rules engine to detect, and

[1] The research is supported by Chinese Natural Science Foundations (50969005,40901262) and by Specialized Research fund of High Education for Inner Mongolia (Njzy08046).

made the related theoretical research work, but did not give a specific implementation strategy [7]. Deal with the problem of missing data values, some academics have suggested the use of inverse interpolation method, the weight-related research and in the field of meteorology has been successfully applied [8].

In this paper, three aspects of the work done. Designed with the characteristics of lake water quality data ETL process. A comprehensive consideration of water quality parameters of water eutrophication factor weighting inverse method interpolation. Designed based on the Drools rule engine kernel dynamic incremental realized abnormal test data volume.

2 Design of the ETL Process with Heterogeneous Data Sources

2.1 Heterogeneous Data Sources

In order to study the issues related to the water eutrophication, data source, on which the analytic environment is established, should consist of two kinds of data: the data of water quality eutrophication and the data representing the key elements of eutrophication. In this paper, the data source mainly includes the following two parts: the water-environment monitoring data and the data manually collected in the Wuliangsuhai Lake. The water-environment monitoring data is collected by the on-line water-quality sensors in the Wuliangsuhai wireless sensor network, and the time granularity of the data sampling is once per 15 minutes, which is conducted in each monitoring station in the Wuliangsuhai Lake. The manual collected data comes from the researchers who collect the related data on the spot in summer, and the associated time granularity is once per day, the sampling space varies each day. Obviously, the data is inconsistent with the manually collected data in terms of time granularity and space granularity.

2.2 Design of the ETL Process

The ETL process provides data for data warehouse. Therefore, the quality of the ETL process relates to the success or failure of the establishment of data warehouse. The ETL processes adopt different strategies and implementation methods for dealing with different data sources. This paper proposed the design of the ETL process which can unify the heterogeneous data sources varying with time granularity and space granularity. The proposed ETL process is as follows.

Firstly, the cleaning of the heterogeneous data is performed with the strategy of data quality controlling introduced next. Secondly, the data uniformity of space granularity is achieved as follows: extract the monitoring data of the sensors located in the sampling stations which are near to the man-made sampling points in terms of latitude and longitude. Following this, the data uniformity of time granularity is achieved as follows: superimposing and fitting of 96 sensor groups of data of 96

sampling times over 24 hours. Thereby, the transformation of data is completed. Finally, the consistent data is loaded in the data warehouse.

3 Data Quality Control

3.1 The Missing or the Abnormal of Water Quality Data

The issues related to the data quality of water-quality sensors mainly are manifested in the two following aspects. One aspect is that WSN is affected by the quality of the communication signal in the data acquisition and transmission. As a result, the sampling data may be missing in some time intervals. The other is that the electrical signal noise and man-made factors may lead to the sampling data abnormal during the process of sensor monitoring. This paper investigates how to deal with the data missing and the data abnormal, which is essential for the control on data quality in the ETL process.

3.2 IWQPW Interpolation

For k points of sensor data in some sampling period employed in the procedure of the data uniformity, the missing and abnormal data may result not only from the parameters of water-quality sensors, but also from the signal strength which may varies with the sampling frequency over one hour. This paper took into account the signal strength and the factors of time period and sampling frequency, and proposed the IWQPW (Inverse Water Quality Parameter Weighting) interpolation method to cope with the issues mentioned above.

Definition1. Assume that the starting instant is t_0, and a group of data is manually collected data over 24 hour period. Accordingly, 96 groups of water-quality sensor data have also been collected, which are indexed by a sequence of integer numbers. Each index corresponds to the related sampling instant.

Definition2. Assume that the data record is a group of manually collected data, the 96 corresponding sequential groups of data is $R_D = \{d_1, d_2, \cdots, d_{96}\}$. For any $d_i (1 \le i \le 96)$ and $d_j (1 \le j \le 96 \ and \ i \ne j)$ in R_D, if their indices are x and y respectively, the sequential distance between two sampling point is calculated as $|x - y|$.

Definition3. Let M (x, y, z, t) be an arbitrary interpolation point in the sample space, N (x_i, y_i, z_i, t_i) be an arbitrary point in 96 sequential sampling values. The sequential weight of the sampling point N, $W_{i,t}$ is defined as:

$$w_{it} = \frac{1}{\sqrt{|l - l_k|}} (1 \le k \le 96) \quad (1)$$

Where l indicates the sequential position of the interpolation point, l_k indicates the sequential position of the sample point.

Definition4. IWQPW (Inverse Water Quality Parameter Weighting) is a spatial temporal sequence interpolation method with the comprehensive consideration of the weight of water quality parameters and the sequential weight. IWQPW takes the distance as weight between the interpolation point and the midpoint of the sample space for weighted average calculation. The sample point is assigned a larger weight if it is nearer to the interpolation point and with a short sequential distance. Assume that a monitoring station is a basis. There are n samples in the 96 sequential sensor sampling space. Let z_i be the value collected of water quality, z be the value of water quality to be estimated as follows:

$$z = \sum_{i=1}^{n} (z_i \times w_i) / \sum_{i=1}^{n} w_i \quad (2)$$

3.3 Dynamic Incremental Rule Engine

3.3.1 Dynamic Incremental Rule Engine Overview

The rule-based engine technology originates from the rule-based expert system. It has been becoming a popular research topic recently that the application of this technology to the ETL process for the control of data quality. Due to the hidden of the abnormal data collected by water quality sensors, it is not easy for non-professionals to tell the abnormal data. This paper proposed to exploit the rule-based engine technology to deal with the issue that how to make the outlier detection rules dynamic increment with the accumulation of expert experiences and the development of related disciplines. However, the following two issues appeared in the application of this technology to the detection of outlier in the sensor spatial data.

- Existing data-cleaning methods based on the rule engine can only process a data record. However, the sensor parameter data is closely related to the data from time and space. Thus, one record data is not sufficient to identify the outliers. Therefore, it is necessary to consider all these related data together.
- It is not flexible of existing cleaning methods based on the rule engine to process the rule file. The historical versions of the domain-expert rule files cannot be made full use. Especially, it is impossible for the existing rules to self-learn and update continuously.

This paper investigated the method for the detection of outliers in the sensor parameter data based on the rule engine technology in order to cope with the above two issues, and proposed a new engine technology of dynamic incremental rules.

3.3.2 The Architecture of the Dynamic Incremental Rule Engine

The designed dynamic incremental rule engine is Drools rules engine as the rules of the nuclear, through plug-in components, to achieve the continuous dynamic growth rule and incremental updates. The architecture of the dynamic incremental rule engine includes:

Rule generation interface. Graphical user interface (GUI) is exploited to edit the user rules and the plan of the rule conversion. In GUI, the rule is represented using custom XML files.

Rule-related job dispatcher. The dispatcher is used to dynamically adjust the priority of the rules and the sequence of the rule conversion.

Rule converter. The rule converter takes the responsibility of the conversion of the custom XML rules to the rules supported by Drools. That is, the converter transfers the object-oriented rules to the Drools-supported rules.

Rule-version controller. With the rule-version controller, the user rules are managed in a centralized manner. The rules with old versions are also regulated. The system maintains a record list regarding the rule usage of each individual user with specific role(s). The personalized interface thus is provided to different users.

Runtime database. The database is used for data persistent storage, which is shared by the other components.

Code generator. The generator produces the code according to the rules, and returns the data anomalies stamp. There are two kinds of return data. One is the PL/SQL code; the other is Java code. The generated code, as input, sends to the module of data layer.

Data layer modules. Data layer modules receive the code, and determine which operation should be performed according to the abnormal stamp. If it is necessary, the interpolation module is called. Then, the persistence operation is performed.

3.3.3 A Sample of the Rule File

This dynamic incremental rule engine designed to use a custom XML file to represent the objects of the rules, call the Java API code embedded in the rules file, rules through five steps to achieve the dynamic and incremental update capability, in detail Process described below.

- Import the classes used in the code, nested the classes in <java:import> tag;
- Describe data set schema. The original data set is divided into the current data set and the other data set, differentiated by the attribute tag, i.e., type. The sample fragment is as follows for the definition of the data set.

```
<java:receive name='sensorDataset'>
 <java:ds  name='currentdataset' type='current'>
   <parameter name='sensor'>
     <class>sensorDataset</class>
   </parameter>
   <parameter name='struct_sensor'  type='srcstruct'>
     <! The structure of data set descriptions >
     <col name='col1'>
       <matched column value='time'></matchedcolumn>
       <datatype value='Date'></datatype>
       <datalength value=7><datalength>
     </col>
     <col name=’col2’> ⋯ </col>
   </parameter>
 </java:ds>
 </java:receive>
```

- Define the function of object operators. The operations are defined between the objects in the set of abstract source data and the objects in the set of training data. The specific function definition is included in the <java:functions> tag. The sample fragment is as follows:

```
<java:functions>
  public int positonInOtherDataset(currentdataset,
           otherdataset, itemtocompare)
  {
    //Return the position of the specified field of the
    //current data set in the other data set ordered by
    //the related values
  }
  </java: functions>
```

- Define the specific rules. The Boolean expressions are used to describe the rule conditions, which consist of the class, the data sets and the operator functions defined in the above, the sample fragments is as follows:

```
<rule-set>
  <rule name="rule1" salience="10">
    <java:condition name1='cond1' cleanitem='phx'>
      isMaxValue (currentdataset, otherdataset, "phx")
        equals the value of a return code
    </java:condition>
  </rule>
  </rule-set>
```

- Define an operation. The described operations are performed when the Boolean expression of the rule condition is true.The sample fragment is as follows:

```
<java:consequence>
  <!Mark the outliers>
    markExceptionData (data sets,outlier row,outlier
    column);
  </java:consequence>
```

3.4 Quality Control Strategy

The process of data cleaning is the most critical part of quality control of water-quality-sensor data, which includes two parts: the interpolation processing of the missing data in water quality monitoring and the detection processing of the outlier. First of all, the missing values of sensor water-quality data are interpolated using IWQPW method. Then, the outlier detection is run using the dynamic incremental rule engine with the input of the processed data set and the rules edited by water experts. Following this, the space interpolation is performed on the outliers using IWQPW method again. Finally, the achieve data set get into the follow-up processing.

4 Experimental Design and Analysis

The purpose of the experiment is to verify the accuracy and reliability of IWQPW interpolation, as well as recall ratio and precision ratio of DIRE rule engine. Recall ratio is calculated as the ratio of the number of detected outliers over the number of statistical outlier. Precision ratio is defined as the ratio of the correct number of detected outliers over the number of detected outliers.

The test set A consists of the four sampling spaces indexed by 1, 2, 3 and 4 respectively. Each sampling space is constructed based on the test sampling space of the real data records of PH, ORP and oxygen content collected in April in the Wuliangsuhai Lake. Then, the test set B is obtained through elimination of the uncompleted record and outliers in the four sampling spaces in the set A under the guidance of related experts.

Experiment 1. Under the guidance of related experts, the statistical number of outliers is achieved manually for each sampling space in the test set A. The number of outliers is also obtained from the DIRE rule engine with the input of each sampling space. Thereby, recall ratio and precision ratio can be calculated. Experimental results are shown in Table 1.

Table 1. Detection of outliers in sensor data with the rule engine

No.	Number outliers	Number detected	Correct number	Recall ratio	Precision ratio
1	364	345	337	94.7%	97.7%
2	253	237	229	93.7%	96.6%
3	377	344	319	91.2%	92.7%
4	423	389	377	92.0%	96.9%
5	340	309	298	90.9%	96.4%

Experiment 2. The purpose of this experiment is to compare the ratio of interpolation accuracy using three following methods. From the test set B, we remove normal data 4 times with the quantities of 105, 150, 245 and 400, respectively. Then, three interpolation methods, i.e., IDW, Kriging and IWQPW are used individually for data interpolation within the allowable range of the error (0.62). The experiment results are shown in Table 2.

Table 2. Accuracy of different interpolation methods for sensor data

Measuring the number of missing values	Interpolation of the average accuracy		
	IDW	Kriging	IWQPW
105	79.2%	77.7%	99.6%
150	77.3%	78.2%	96.2%
245	66.0%	66.9%	95.4%
400	70.3%	70.5%	90.0%

The experiment results show to some extent that our proposed method is useful in practice and effective. The results also show that our methods strongly rely on the rules.

5 Conclusion

The objective of this paper is to deal with the issues related to the quality control of water-quality sensor data. The ETL process is first proposed in order to establish data warehouse. In addition, the quality control strategy is proposed which combines IWQPW method with the dynamic incremental rule engine. Thereby, it can be clean up that the missing values and outliers. The future work is to investigate further universal interpolation methods and the self-learning capability of the rule engine.

6 References

1. Chen, Q., Mynett, A.E.: Integration of data mining techniques and heuristic knowledge in fuzzy logic modelling of eutrophication in Taihu Lake. J. Ecological Modelling. 162, 55--67 (2003)
2. Lenat, D.R.: Water Quality Assessment of Streams Using a Qualitative Collection Method for Benthic Macroinvertebrates. J. Journal of the North American Benthological Society. 7, 222--233 (1998)
3. Neal C., Robson, A.J.: A summary of river water quality data collected within the Land-Ocean Interaction Study: Core data for eastern UK rivers draining to the North Sea. J. Science of the Total Environmen. 251, 585--665 (2000)
4. Codd, G.A.: Cyanobacterial toxins, the perception of water quality, and the prioritisation of eutrophication control. J. Ecological Engineering. 16, 51--60 (2000)
5. Vassiliadis, P., Simitsis, A., Skiadopoulos, S.: Conceptual modeling for ETL processes. In: 5th ACM international workshop on Data Warehousing and OLAP, pp. 14--21. ACM, New York (2002)
6. Simitsis, A.: Mapping conceptual to logical models for ETL processes. In: 8th ACM international workshop on Data warehousing and OLAP, pp. 67--76. ACM, New York (2005)
7. Loshin, D.: Rule-based data quality. In: Proceedings of the eleventh international conference on Information and knowledge management, pp. 614--616. ACM, New York (2002)
8. Sun, Y., Kang, S., Li, F.,Zhang, L.: Comparison of interpolation methods for depth to groundwater and its temporal and spatial variations in the Minqin oasis of northwest China. J. Environmental Modeling & Software. 24, 1163--1170 (2009)