ETL Process Design and Quality Control Research Based on Radar Spatial Data

Xuejun Chen[1, 2]

[1]Key Laboratory of Western China's Environmental Systems(Ministry of Education)&College of Earth and Environment Sciences, Lanzhou University, Lanzhou 730000, P.R. China
[2]Gansu Province Meteorological Information Center, Lanzhou 730020, P.R. China

Abstract

In order to study nowcasting operations using spatial data mining, a spatial data warehouse must be built via ETL process to store radar spatial data reflecting strong convection weather. Also, the quality control of radar spatial data in the ETL process, which directed at weather station observed data and radar spatial data, based on dynamic incremental rule engine(DIRE) and ISTDW method. The experimental result demonstrates that DIRE can effectively uncover the abnormal value in radar data and that ISTDW can estimate the absent values in radar data via interpolation with the tolerance of errors.

Keywords: Radar Spatial Data, ETL, DIRE, ISTDW

1. Introduction

Up to now, some researchers both Domestic and international have done some research in relation to integrated analysis and process of complicated meteorological operations via spatial data mining[1]-[3], especially nowcasting based on multiple isomeric data sources[4]. And the facts show that building a data warehouse for strong convectional meteorological data is the foundation of nowcasting operations and the design of ETL process plays a vital role in constructing an effective analysis platform.

Previous research focuses on the conceptual model of ETL process and the transformation from conceptual model to logical model of ETL process， and provides a basis for the construction of ETL model[5]-[6]. Some paper poses a rule-based engine to detect the abnormal data and discusses the design of rule engine from the sight of theory for the data quality control of the process of ETL[7]. Some also provide an improved method of IDW (Inverse Distance Weighting), which consider the different changes of meteorological factor in longitude and latitude but pay no attention to the effect of time and height[8].

The paper is organized as follows: the second section shows the ETL process; the third section represents the control of data quality and the fourth design and analysis of experiment, the last section is a conclusion.

2. ETL Process on Isomeric Data Sources

2.1. Isomeric Data Sources

Nowcasting is mainly affected by the strong convectional weather condition, thus, the data source for constructing an analysis platform should consist of realtime weather data and convectional weather data. The data source this paper involved consist mainly of radar spatial database and weather station observing files(A-files) , database mainly stores data collected by radar in last few years with the time granularity 1/6 times per minute and the space granularity covered by radar scanning range; A-files store the observational data collected by weather stations and realtime data with time granularity 1 times per hour and space granularity a point determined by the specified longitude and latitude. So, there's inconsistence of time and space granularity between radar data and station data.

2.2. Design of ETL Process

ETL provides data for data warehouse, so its design determines the possible analysis result of a data warehouse. Different policies and implemental methods are selected to deal with different data sources; this paper designs an ETL process to ensure the consistence of time and space granularity between isomeric data sources. First of all, clean the data according to data quality controlling policies represented by this paper, then, extracting the radar scanning points whose longitude and latitude are so close to stations' according to the longitude and latitude of weather stations to realize the space granularity corresponding between stations' and radars'; consequently, realize the time granularity consistence via superposition and fitting on 10 groups radar data in an hour; finally, load the result data to

data warehouse. The logic of these processs can be demonstrated by Fig.1.

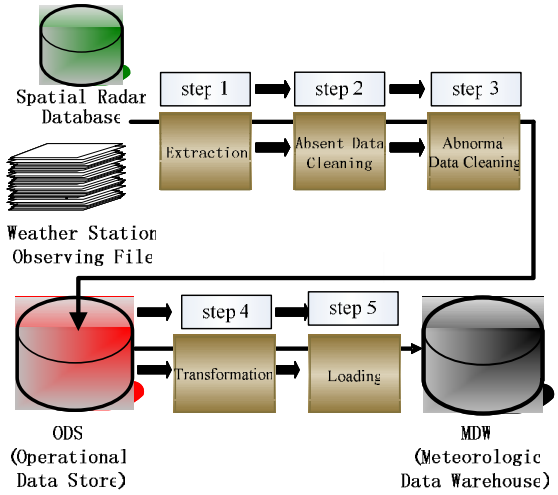

Fig. 1: Logical Structure of ETL.

3. Data Quality Control

3.1. The Absent and Abnormal of Radar Data

There are two factors that will affect the quality of radar data: 1. features of ground surface will cause the lost data in some positions when radar is scanning; 2. electrical noises and artificial factors will drive the data abnormal. How to resolve these problems is so vital and will be discussed deeply in this paper.

3.2. ISTDW Interpolation Method

In the process maintaining the consistence of isomeric data sources, the absent values or abnormal values of k points in radar scanning space are not only related to space distance such as longitude, latitude, height etc. but also affected by scanning sequences in an hour. This paper considers the influences of longitude, latitude, scanning height in different elevations and scanning series synthetically, advances a interpolation method, namely, ISTDW (Inverse Spatial Time Distance Weighting).

Difinition1.Supposed from time t0, a group of station observing data and the corresponding 10 groups of radar scanning data are obtained, label the 10 time series from time t_0 to t_0+10 with integer numbers in succession, then the number of each time series is defined as the time series position.

Difinition2.Given data records E_D are a group of station observing data, the corresponding 10 time series data are $R_D = \{d_1, d_2, \ , d_{10}\}$,for any $d_i (1 \le i \le 10)$ and $d_j (1 \le j \le 10, j \ne i)$ in R_D,if the time series position of d_i and d_j are x and y respectively, then $|x - y|$ is defined as the time series distance of points d_i and d_j.

Difinition3. Given M(x, y, z, t) is any interpolating point in radar scanning space and also is a point in 4-dimentional space spaned by longitude, latitude, height and time. $N(x_i,y_i,z_i,t_i)$ is any sample point in corresponding 10 time series, then the time series weight of N is defined as:

$w_{it} = \frac{1}{\sqrt{|l - l_k|}} (1 \le k \le 10)$, where l is the time series position of M, l_k indicates the time every position of N.

Difinition4.Given M(x, y, z, t) is any interpolating point in radar scanning space and also is a point in 4-dimentional space spaned by longitude, latitude, height and time. $N(x_i,y_i,z_i,t_i)$ is any sample point in corresponding 10 time series, then the longitude weight, latitude weight, height weight of N

$$w_{i,x} = \frac{x - x_i}{d_{i,x} \times d_{i,x} + d_{i,y} \times d_{i,y} + d_{i,z} \times d_{i,z}}$$

$$w_{i,y} = \frac{y - y_i}{d_{i,x} \times d_{i,x} + d_{i,y} \times d_{i,y} + d_{i,z} \times d_{i,z}}$$

$$w_{i,z} = \frac{h - h_i}{d_{i,x} \times d_{i,x} + d_{i,y} \times d_{i,y} + d_{i,z} \times d_{i,z}}$$

are defined respectively as $w_{i,x}$, $w_{i,y}$, $w_{i,z}$:

So, from Difinition3 and 4, we can get the weight w_i of (x_i, y_i, z_i, w_i), the calculation is demonstrated by the following formulas:

$$s = \sum_{i=1}^{n} \frac{1}{d_{i,x} \times d_{i,x} + d_{i,y} \times d_{i,y} + d_{i,z} \times d_{i,z}}$$

$$s_x = \sum_{i=1}^{n} w_{i,x} \qquad s_y = \sum_{i=1}^{n} w_{i,y}$$

$$s_z = \sum_{i=1}^{n} w_{i,z} \qquad s_t = \sum_{i=1}^{n} w_{i,t}$$

$$m_{i,x} = s - w_{i,x} \times (s_x - w_{i,x})$$

$$m_{i,y} = w_{i,y} \times (s_y - w_{i,y})$$

$$m_{i,z} = w_{i,z} \times (s_z - w_{i,z})$$

$$m_{i,t} = w_{i,t}(s_t - w_{i,t})$$

$$w_i = \frac{\sum_{i=1}^{n}[m_{i,x} - m_{i,y} - m_{i,z} + m_{i,t}]}{d_{i,x} \times d_{i,x} + d_{i,y} \times d_{i,y} + d_{i,z} \times d_{i,z}}$$

ISTDW (Inverse Spatial Time Distance Weighting) is a space-time series interpolating method which considers synthetically the longitude weight, latitude weight, height weight and time series weight, it uses the distance between interpolating point and any point in sample space as the weight to calculate the average weight, the closer the sample point is away from the interpolating point and the smaller the time every distance is, the greater the weight will be assigned. Given a specified station as the reference and n sample points taken from 10 radar cylinder spaces, zi denotes the real meteorologic value while z the estimating value to be calculated by interpolation, then z satisfies the following formula:

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

3.3. Dynamic Incremental Rule Engine DIRE

3.3.1. Introduction to DIRE

The rule engine method is originated from rule based specialist system. Applying it to a ETL process appeared in recent years and soon becoming a research hotspot.Usually,radar scaned data remain in concealment and hard to be discovered by non specialist.The question lies in how to make the detecting of abnormal value of radar scaned data stick out a mile by the knowledge accumulation of specialist in some areas or the development of the subject.This paper introduced rule engine method to solve the problem.

The following two questions appeared when use this method to detect abnormal value of radar spatial data.

- Firstly, one record is detected every time by the previous product of rule engine.But the things changed enough when apply it to radar spatial data, which is strongly correlated to time and space.Difficult to distinguish which value is abnormal only considering one record and don't take into account other record which may play a more important role.So, time and space feature of radar spatial data should be synthetically considered.
- Secondly, the flexibility of the previous method is not enough.This question mostly lies in history version of rule established by authority on the area can only be use a few short time.Especially lies in the feature of self-adaptive or self-study is not realized .

In the paper, we bring forward a solution on the above two problems.which is called Dynamic Incremental Rule Engine (DIRE).

3.3.2. The Architecture of DIRE

The core component of DIRE is Drools' .Some external components are added to the core to expand the function of Drools' impletementation.In brief,the rules in DIRE can dynamic enhance itself and update incrementally.In addition, it can represent more complex rules,especially object-base rules. The Fig.2 represent its architecture.

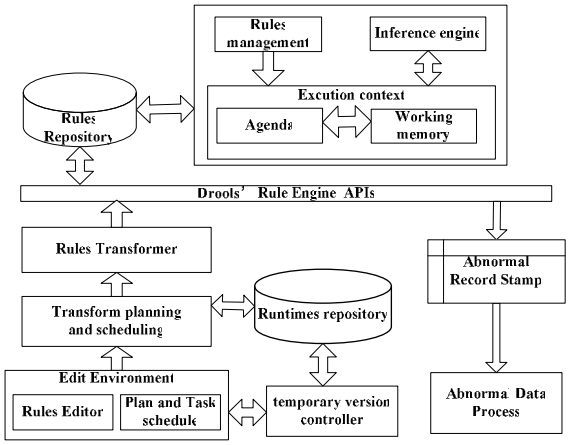

Fig.2 : Architecture of DIRE.

In Fig.2, every component is described in detail as follows.

- **Rule Edit Environment** The environment provides an approach for meteorologists to edit their rules or rule transforming plan.This is a GUI interface which offered a rich edit function.So, user can represent their real mind through this interface. The rule refered will be discussed particularly in next section. In fact, it is a XML file with special structure.
- **Schedule and Plan of Rule Transformation** The rule which many objects is embedded, such as JAVA class, function. should be translated into a comprehensible representation. The component is designed to ease user to fully master the process of rule

transformation. Users can design a transformation plan and schedule it in their own way.

- **Rule Transformer** This is a real translation component for rule transformation.The execution result of the component has changed the represent of user defined rule to the rule which can be understood by Drools rule engine.
- **Version and User Controlor** The design of this component thinks over the personality of different user, who may have deferent mind. Every user has a indepdent enviroment belonging to himself. In addition, the history version of a rule may be useful for users to enhance their design. The system provides a history version for each user with certain roles.
- **Runtime Repository** The runtime repository is a permanent storage for system data and metadata. Rules, rule execution plan, users' project, users' role, etc. are stored in this place and the other component can read or write the data stored in the infrastructure.
- **Code Generator** Although the Fig.2 doesn't describe this component, the component really exists in our system and plays an important role in each data process procedure. The component is used to generate code and the code generated may be inserted into our application to access data.
- **Data Component** This is a fundamental component which provides a channel between our application and database.

3.3.3. An example for rule file

The dynamic incremental rule engine designed and implemented in this paper use a elaborate XML file describe objects and operation of the objects. Mang JAVA objects are embeded in the rule and are devided into five sections. Each of section is described in detail as follows.

Import the Classes Used in the Code As shown below, the statement embeded into a pair of <java:import> tags.

Description of Dataset Structure Devide original dataset into two parts: current dataset and other dataset.A attribe type used to distinguish this two parts.The code segment show as follows.

```
<java:input name='radarDataset'>
  <java:dataset name='currentdataset'
  type='current'>
    <parameter name='radar'>
      <class>radarDataset</class>
    </parameter>
    <parameter name='structure_radar'
    type='sourcestruct'>
      <!describe the structure of dataset>
      <col name='col1'>
      <matchedcolumn
      value='time'></matchedcolumn >
      <datatype value='Date' ></datatype>
      <datalength value=19><datalength>
      </col>
      <col name='col2'>…</col>
    </parameter>…
  </java:dataset>
</java:input>
```

Definition of Object Operator Use object operator to describe the relationship of source dataset and other dataset.the Difinition of operator function look like follows.

```
<java: functions>
  public int
  positonInOtherDataset(currentDataset,otherDatas
  et,itemToCompare)
  {
    //return the order of certaion field of current
    data set in the other dataset.
    //usually,the size of current dataset is larger
    than the size of other data set.
  }
</java: functions>
```

Definition and Representation of Rule The section discribes the rule active condition with the things listed in section above. In fact, the conditions is a series of bool expressions.The code segment lies in follows.

```
<rule-set>
  <rule name='rule1' salience='10'>
    <java: condition name1='condition1'
    cleanitem='dbz'>
    isMaxValue (currentdataset, otherdataset, 'dbz')
    == certain return code
    </java: condition>
  </rule>
</ rule-set >
```

Difinition of Conseqence While the Rule Actived The section may be a place to set the consequence of the rule actived due to the result of rule conditions discussed in previous section.we list a simple example as below.

```
<java: consequence>
  <! mark up the position of abnormal value>
  markExceptionData(currentDataset,the row
  number of abnormal value lies in，the column
  number of abnormal value lies in);
</java:consequence>
```

3.4. Policies of Data Quality Control

The key part of radar quality control is data cleanning of ETL process. This contains two parts:spatial interpolation for absent value and detection of abnormal value. These two tasks have been fulfilled follow hard after the step listed here: Firstly, The absent value of radar spatial data is processed using the ISTDW method discussed above. Through the step, a dataset without absent value can be achieved. Secondly, the dataset obtained from previous step and the rule edited by meteorologist become a steady input of the DIRE rule engine. In this infrastructure, the abnormal value of radar is detected and marked up. Thirdly, the output of previous step with mang abnormal stamp becomes an input of the mothod described firstly, the abnormal data with abnormal stamp is replaced by the value of spitail interpolation with ISTDW method. After the process, the dataset can be considered as clean and the next processes, such as transforming, loading can be get along well-off.The whole process is described as following Fig.3.

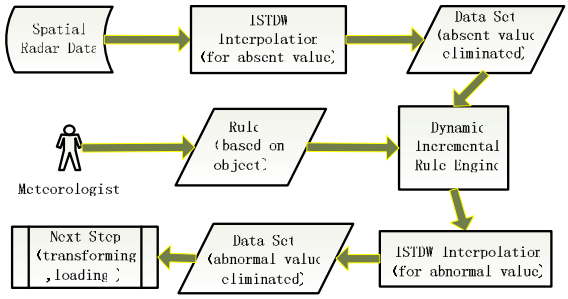

Fig.3: Policies of Data Quality Control.

4. Experiments design and analysis

4.1. Construction of testing dataset

Two testing datasets are constructed: set A and set B, for the sake of accuracy, credibility of ISDW method and precision and recall of DIRE rule engine. More than 3500 million radar scaned data, which is observed by the radar lies in Lanzhou city, Gansu province from May to September in 2006, become the population. Each of sample spaces is extracted from the population and represents radar data in certain hour in the population.In the experiments, 5 sample spaces are used in all and listed in following table.

4.2. Result of Experiment

Test 1. In the test, Data set A is choosed as target data set. Marking up abnormal data under the instruction of meteorologist and input this dataset to our DIRE rule engine.Then, two measures for the rule engine have been difined. One called completeness, a ratio of the number of abnormal value detected to the number of abnormal value marked by meteorologist. The other measure, called veracity, is a ratio of the number of truely abnormal value detected to all the number of abnormal detected.The result of the test show as table2.

No.	Observed Time	Sample Space Size
1	5/15/2006 18:00	8908
2	6/6/2006 14:00	9817
3	7/22/2006 4:00	10190
4	8/9/2006 15:00	9305
5	9/10/2006 20:00	9779

Table1: Size of Sample Space(Entry).

4.3. Result of Experiment

Test 1. In the test, Data set A is choosed as target data set. Marking up abnormal data under the instruction of meteorologist and input this dataset to our DIRE rule engine.Then, two measures for the rule engine have been difined. One called completeness, a ratio of the number of abnormal value detected to the number of abnormal value marked by meteorologist. The other measure, called veracity, is a ratio of the number of truely abnormal value detected to all the number of abnormal detected.The result of the test show as table2.

No.	NTAV	NAVD	NAVTD	Precision	Recall
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%

(NTAV is Number of Truly Abnormal Value, NAVD is Number of Abnormal value Detected, NAVTD is Number of Abnormal Value Truly Detected)

Table 2: Result of DIRE rule engine for abnormal value detecting.

been difined. One called completeness, a ratio of the number of abnormal value detected to the number of abnormal value marked by meteorologist. The other measure, called veracity, is a ratio of the number of truely abnormal value detected to all the number of abnormal detected.The result of the test show as table2.

As shown in table 2, the precision and the recall of our DIRE rule engine applying to abnormal value detecting both obtained a high performance, more than 90%.Only few sample spaces with unconspicuous feature of strong meteorologic convection appeared errors or non-detected cases.But, mang aggregation operation for radar data will be carry through to obtain a consistent granularity with the observed data of

weather station.So the effect of these errors or non_detected cases may be ignorable.

Test 2. In the test, dataset B is choosed as our target dataset. Some normal value are eliminated at four times, respective 105,150,245,400 and three kinds of methods, IDW 、 Kriging 、 ISTDW, used to interpolate value to test the ISDW methods. The effect of three kind methods is shown in following table with the tolerance of certaion error range.

Number of Absent Value	Average Precision		
	IDW	Kriging	ISTDW
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%

Table 3: Accuracy of Three Kind of Methods.

As shown in table 3, the accuracy of ISTDW is clearly higher than the other two methods. In addition, the stability of ISTDW is enhanced with the increasement of the number of absent value.

5. Conclusions

In the paper, an ETL process is designed to gain a consistency granularity of time and space and to improve the data quality. A DIRE rule engine and an ISTDW spatial interpolation method is held out to realize radar data cleaning. As shown in above two tests, the accuracy of ISTDW is clearly higher than the other two methods, IDW and Kriging. The precision and the recall of the DIRE rule engine applying to abnormal value detecting both obtained a high performance, more than 90%. So, the meteorologic data warehouse, which is implemented by the ETL process, will provide a massiness data platform for nowcasting operations.

Acknowledgement

Under the auspices of the Ministry of Science and Technology Basic Development Program of China(No. 2005DKA31700).

References

[1] J.T. Anthony Lee, R.W. Hony, W.M. Ko,etc. Mining spatial association rules in image databases.Information Sciences, *an Internation al Journal*, 177 (7):1593 –1608, 2007.

[2] D. Birant, A. Kut, ST-DBSCAN: An algorithm for clustering spatial-temporal data. *Data & Knowledge Engineering* , 60(1):208-221, 2007.

[3] Y.B. Yang, H. Lin, G.Z. Yang, *A meteorol ogyical conceptual modeling approach based on spatial data mining and knowledge discovery. Innovations in Applied Artificial Intelligence* , Bari, Italy: Springer-Verlag GmbH, pp.490 – 499, 2005.

[4] A. Sarkka, E. Renshaw. *The analysis of marked point patterns evolving throuth space and time*, 51(3): 1698 – 1718, 2006.

[5] P. Vassiliadis, A. Simitsis, S. S. oulos. Conceptual modeling for ETL processe s, *Proceedings of the 5th ACM international workshop on Data Warehousing and OLAP*, 2002.

[6] A. Simitsis, Mapping conceptual to logical mo dels for ETL processes, *Proceedings of the 8th ACM international workshop on Data warehousing and OLAP*, pp.67-76, 2005.

[7] D. Loshin, Rule based data quality, *Proceedings of the eleventh internatio nal conference on Information and knowledge management*, pp.614-616, 2002.

[8] Zh.H. Lin, X.G. Xia, H.X. Li, etc. Comparison of Three Spatial Interpolation Methods for Climate Variables in China, *Geography Journal*, 57(1):27-56, 2002.