Big data 3-1

1 Hello everyone, I am Haiying Che, from Institute of Data Science and knowledge Engineering

School of Computer Science, in Beijing Institute of Technology, in this session, we will discuss Data Preprocessing.

2 Data Preprocessing mainly includes data cleaning Data transformation and data reduction

3

Why do we need to do data cleaning? Because when we collect data from different data resources, we inevitably have some dirty data in the results.

Let’s look at an example, in this table, in the second row , for the record id=2, the age attributes is missing.

If we calculate the average age, this missing age value will Serious impact the average result.

and the third and fourth records are repeated. And the sixth record is conflict record, according to the age and the birth data.

the seventh record is abnormal age 101 according to the birth data. All these dirty data will damage the analysis results.

4

According to Murphy's Law: As long as everything **can** go wrong, it **will** go wrong.

5

How to prevent dirty data from appearing, we can try 2 different ways.

1) Develop data standards:

* Unify attribute value encoding of multiple data sources
* Give the attribute name and attribute value as clear as possible

2) Optimize system design:

* Use options as much as possible for key attributes instead of manually filling in the entry.
* Important attributes appear in a prominent position, use required options
* Outliers should be modified

6

Data preprocessing could include data cleaning; data transformation and data reduction

1) data cleaning includes missing data processing and noisy data processing.

1.1 For missing data we can directly ignore the tuple, or record. Or we can fill the missing values manually, by mean or by most probable value.

1.2 eliminating the noisy data methods include binning methods, regression and clustering.

2) As for data transformation, it includes Normalization, attributes selection, discretization and concept hierarchy generation.

3) The third aspect of data preprocessing is data reduction.

Data reduction includes data cube aggregation, attribute subset selection, numerosity reduction which is Reduce the amount of data and dimensionality reduction.

**4) Data integration:**

Store all data in a database, data warehouse or file to form a complete data set.

**5) Data specification:**

Eliminate the data attributes that cannot describe the key characteristics of the system, and only retain part of the data attribute set that can describe the key characteristics.

And others…

7

The research on data cleaning technology first began with the correction of the US Social Security number. Later, with the rapid development of information and commerce,

the research in this area was accelerated, including

1) Perform anomaly detection on the data set,

usually using statistical methods to detect the numerical attributes of the data. By calculating the mean and standard deviation of the attribute values ​​and other indicators, identify the anomalous attributes and records within the confidence interval of each attribute.

2) Deduplication of data:

The process of data deduplication is to repeat the process of cleaning data records.

This process is particularly important in data warehouse applications, because when integrating data from different data sources, many duplicate data records may be generated.

3) Cleaning of missing data: Most of them use approximate values ​​to replace missing values ​​to clean the data.

Methods to obtain approximate values ​​include Bayesian networks, neural networks, KNN classification, rough set theory, etc.

The core of these methods is to judge missing records and Similarity between other complete records.

Domestic research in this area is still in its initial stage.

Data cleaning is mainly concentrated in industries that require high accuracy of customer data, such as banking, insurance, and securities. These industries only do data cleaning for their customers, and only develop software for specific applications, without general products

8

It takes a long time to analyze and mine massive data. To make data mining more effective, data needs to be regulated. The main research content of the data reduction:

1) Dimensionality reduction processing of high-**dimension** data

The process mainly adopts the method of market redundant data attributes, and deleting redundant data attributes requires business knowledge in a certain field. Commonly used dimensionality reduction methods include stepwise forward selection method, stepwise backward deletion method, and decision tree induction method Wait

2) Reduce the amount of data

When it takes a long time to process a large amount of data, it cannot meet the requirements of some applications with high real-time requirements. At this time, the amount of data needs to be reduced. The methods of comrades in this process, including histogram, clustering, etc., and then select smaller-scale data from the data set

3) Data Discretization Technology

This technology can convert continuous attributes into discrete attribute values, reducing the number of attribute values, thereby reducing the computing time for processing data

9

**In this session we general introduced** Data Preprocessing.

**thank you for your attention, if you have any question, feel free to contact me.**