Big data analysis 3-4

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 transform .

2

**Data transform can include Data integration and Data transform. Let’s look at Data integration first.**

3

**Data integration:**

Integrate data from multiple data sources into a consistent storage, which could involve the task Pattern matching, Data redundancy processing and Data value conflict solving;

Let’s look these 3 tasks one by one.

4

First one is Pattern matching. Integrate metadata from different data sources.

We can recognize the real-world entity from different data sources and mapping them together,

like A.cust-id=B.customer\_no.

5

Then deal with Data redundancy issue,

* The same attribute will have different field names in different databases.
* One attribute can be derived from another attribute. For example, the average monthly income attribute in a customer data table can be calculated based on the monthly income attribute.

And Some redundancy can be detected by correlation analysis.

6

Then solve the Data value conflict problem.

For a real-world entity, its attribute values from different data sources may be different.

Such as Differences in representation, different scales, or differences in coding, etc.

For example:

**the weight attribute uses the metric system, like kg, g** in one system, but **uses the imperial system like pound** in another system.

Same price attributes in different locations using different currency units, $, pound, RMB

7

After data integration , then let’ come to data transform.

8

Data transform means, in order to facilitate efficient analysis , we need change the data from one form to another form.

We can use smooth methods like Binning Clustering Regression to eliminate the noise Or discretize continuous data, and increase granularity.

By doing this, we can reduce the data amount for further analysis.

9

Second transformation is Aggregation.

By doing avg(), count(), sum(), min(), max()...calculation,

we can use calculation result to represent the detail data.

For example: daily sales (data) can be aggregated to get the monthly or annual total.

And these aggregation data are more convenient for big data analysis.

10

Third transformation is Data generalization. Which means **Replace low-level data objects with more abstract (higher-level) concepts.**

For example:

street attributes can be generalized to higher-level concepts, such as: city, country.

Similarly, numeric attributes, such as age attributes, can be mapped to higher-level concepts, such as young, middle-aged, and old.

11

The fourth transformation is Data Normalization.

Which means The data is scaled proportionally to make it fall into a specific area, so as to eliminate the deviation of the mining results caused by the different sizes of the numerical attributes.

Such as mapping the salary income attribute value to the range of [-1.0,1.0].

The methods for Data Normalization could be

(1) Min-Max normalization

(2) Zero-mean normalization (z-score normalization)

(3) Standardization of decimal calibration

12

The fifth transformation is Attribute construction.

Use the existing attribute set to construct new attributes and add them to the existing attribute set to help dig deeper pattern knowledge and improve the accuracy of mining results.

For example: According to the width and height attributes, a new attribute can be constructed: area.

13

In this session we discussed Data integration and data transform.

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