Big data 3-2

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 Quality

2

About data quality, we want to discuss, **Data quality of Single Data resource and Data quality of Multiple Data resource**

3

The classification of data quality issues has two dimensions,

1) Classified according to the number of data sources, data quality problems can be divided into two types: single data source and multiple data sources

2 )Classified according to abstract level or concrete level. Data quality issues can be divided into model-level issues and instance-level issues.

Model level problems can be solved by improving model design, model conversion and model integration.

instance level Data quality problems refer to errors and inconsistencies in the actual data content.

These problems are often invisible at the model level and need to be resolved in the data cleaning process.

For the Single data resource data quality problems , in model level it could be Uniqueness, Integrity constraints, Attribute dependence and so on.

For the Single data resource data quality problems , in instance level it could be Spelling errors, Redundant and repeated records, Attribute value conflict and so on.

For the Multiple data resources data quality problems , in model level it could be Naming conflict, Structure conflict and so on.

For the Multiple data resources data quality problems , in instance level it could be Different representations of values, Overlapping/duplicate records and so on.

4

Let ‘s look at the **Single Data resource Model level and Instance level data quality issues first**

5

For a single data source, the data quality problem at the model layer largely depends on the integrity constraints of the design model on the data.

The integrity constraints of the database determine which data values are acceptable.

For example, a data to represent a date, the format and type of the date need to be restricted to ensure that the format of all date data in the database is uniform.

However, for data sources that do not have uniform constraints such as files and web data, the possibility of causing data value errors and inconsistencies is greatly increased.

6

For a single data source, data quality problems at the instance level are unavoidable at the model design level, such as data input errors, etc.

7

Single Data resource data quality errors categories include

1) Attribute internal: limited to a single attribute value, such as age: 2000

2) Inside the record (between attributes): Different attribute values in the same record are inconsistent, for example, age and birthday cannot correspond

3 )Internal data source (between records): inconsistency between different records of the same data source, for example, the name of the same ID is inconsistent

4 )Between data sources: Some attributes in the data source are inconsistent with related values in other data sources, for example, the age corresponding to the same ID is inconsistent

The data quality problems of different levels and scopes correspond to different data cleaning methods. A clear data level range is the basis for finding a reasonable data cleaning method

8

Now Let ‘s look at the **Multiple Data resource Model level and Instance level data quality issues**

9

In the case of multiple data sources, data from different data sources need to be integrated. Each data source is often created by a specific application to meet the needs of specific users, and the data model design of each data source will be very different. In addition, each data source may contain dirty data, and different data sources may have different representations, data duplication, or data conflicts for the same data.

Therefore, the data quality problem in the case of a single data source is in the case of multiple data sources The next still exists. In addition, in the case of multiple data sources, data cleaning faces many new problems, such as structural conflicts, naming conflicts, duplicate records, etc.

The main problems at the multi-data source model level are naming conflicts and structural conflicts:

Naming conflict refers to using the same name for different data objects or using different names for the same data object.

There are many different situations of structural conflicts, which usually refer to different ways to represent the same data object in different data sources. For example, the same object has different attribute granularity in different data sets, different composition structures, different data types, and different Integrity constraints, etc.

10

The conflict at the data instance level refers to the conflict of specific data.

The data quality problems existing in a single data source may appear in different forms in different data sources, such as record duplication and record conflict. Even if different data sources have the same attribute name and data type, there may be different data value representations. For example, the description of gender can be expressed as male, female, or as M, F.

Or the data deserves different interpretations, such as the U.S. dollar, euro and other currency measurement units

In addition, the information provided by different data sources may be aggregated at different levels. For example, a single record in a data source describes the sales information of a certain product, while a record in another data source describes a group of similar products. sales information

11

**In this session we discussed some concrete** Data quality issues.

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