Data Preprocessing

——Why pre-process data?——

徐华

清华大学 计算机系 智能技术与系统国家重点实验室

xuhua@tsinghua.edu.cn

1

Data Preprocessing

- About data
- Why preprocess the data?
- Descriptive data summarization
- Data cleaning
- Data integration and transformation
- Data reduction
- Discretization and concept hierarchy generation
- Summary

2

Why Data Preprocessing(1)

- **Data in the real world is dirty.**
  - **Incomplete(不完整): lacking attribute values, lacking certain attributes of interest, or containing only aggregate data**
    - **e.g., occupation= “”**
  - **Noisy (有噪音) : containing errors or outliers**
    - **e.g., Salary= “-10”**
  - **Inconsistent (不一致) : containing discrepancies (冲突) in codes or names:**
    - **External discrepancies**
      - **e.g., Age= “42” Birthday= “03/07/1997”**
      - **e.g., Was rating “1,2,3”, now rating “A, B, C”**
      - **e.g., discrepancy between duplicate records**
    - **Internal discrepancies**
      - **e.g., IngrA(10)+IngrB(3)+IngrC(4) -> Germ(70%)**
      - **IngrA(13)+IngrB(2)+IngrC(4) -> Germ(65%)**

3

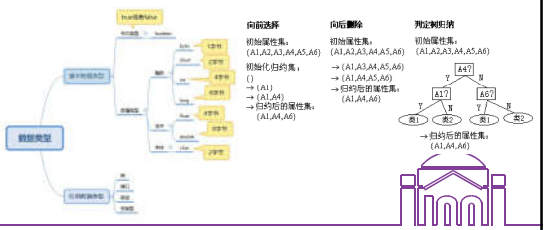
Why Data Preprocessing(2)

- **Incomplete data comes from**
  - **Different consideration between the time when the data was collected and when it is analyzed.**
  - **Human/hardware/software problems**
- **Noisy data comes from the process of data**
  - **Collection**
  - **Entry**
  - **Transmission**
  - **Conflict with common sense**
- **Inconsistent data comes from**
  - **Different data sources** (Web , Manual Collections, Special Equipments, Database)
  - **Actual Experiment Equipments** (Sensors)
  - **Different Environment Conditions** (IOT Equipments)

4

Why Data Preprocessing(3)

- The form of data/information needs to be transformed.
  - Different data need to be integrated.
    - e.g. In Table A: Age = “” ;
    - e.g. In Table B: Weight= “”
  - Different data need to be transformed.
    - e.g. questionnaire data
  - Different data need to be discretized.
  - Different data need to be reduced.



5

Why Is Data Preprocessing Important?

- No quality data, no quality mining results!
  - Quality decisions must be based on quality data
    - e.g., duplicate or missing data may cause incorrect or even misleading statistics.
  - Data warehouse needs consistent integration of quality data
- Data extraction, cleaning, and transformation comprises the majority of the work of building a data warehouse

6

Multi-Dimensional Measure of Data Quality

- A well-accepted multidimensional view:
  - Accuracy (准确的)
  - Completeness (完整的)
  - Consistency (一致的)
  - Timelines (合时的)
  - Believability (可信的)
  - Value added (有附加价值的)
  - Interpretability (可解释的)
  - Accessibility (可存取的)
- Broad categories:
  - Intrinsic(本质的), contextual (相关的) , representational (代表性的) , and accessibility (可存取的) .

7

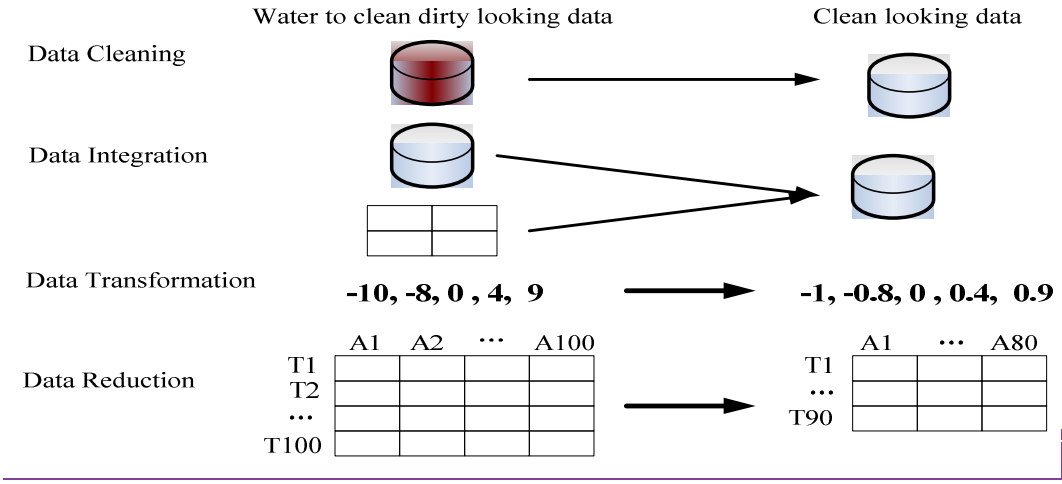
Major Tasks in Data Preprocessing

- Data cleaning
  - Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies
- Data integration
  - Integration of multiple databases, data cubes, or files
- Data transformation
  - Normalization and aggregation
- Data reduction
  - Obtains reduced representation in volume but produces the same or similar analytical results
- Data discretization
  - Part of data reduction but with particular importance, especially for numerical data

8

Form of data preprocess

- **Key Steps: Grasp and understand the data.**



9

Thanks !

10