Data Mining

2.1 Data Preprocessing

Fall 2008

Instructor: Dr. Masoud Yaghini

Outline

- Why Data Preprocessing?
- Major Tasks in Data Preprocessing
- References

Why Data Preprocessing?

Why Data Preprocessing?

- Data in the real world is dirty
  - incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
    - e.g., occupation=“ ”, martial status =“ ”
  - noisy: containing errors or outliers
    - e.g., Salary=“-10”
  - inconsistent: containing inconsistencies in codes or names
    - e.g., Age=“42” Birthday=“03/07/1997”
    - e.g., Was rating “1,2,3”, now rating “A, B, C”
    - e.g., inconsistency between duplicate records

Why Is Data Dirty?

- Incomplete data may come from
  - "Not applicable" data value when collected
  - Different considerations between the time when the data was collected and when it is analyzed.
  - Human/hardware/software problems
- Noisy data (incorrect values) may come from
  - Faulty data collection instruments
  - Human or computer error at data entry
  - Errors in data transmission
- Inconsistent data may come from
  - Different data sources
  - Functional dependency violation (e.g., modify some linked data)

Why Is Data Preprocessing Important?

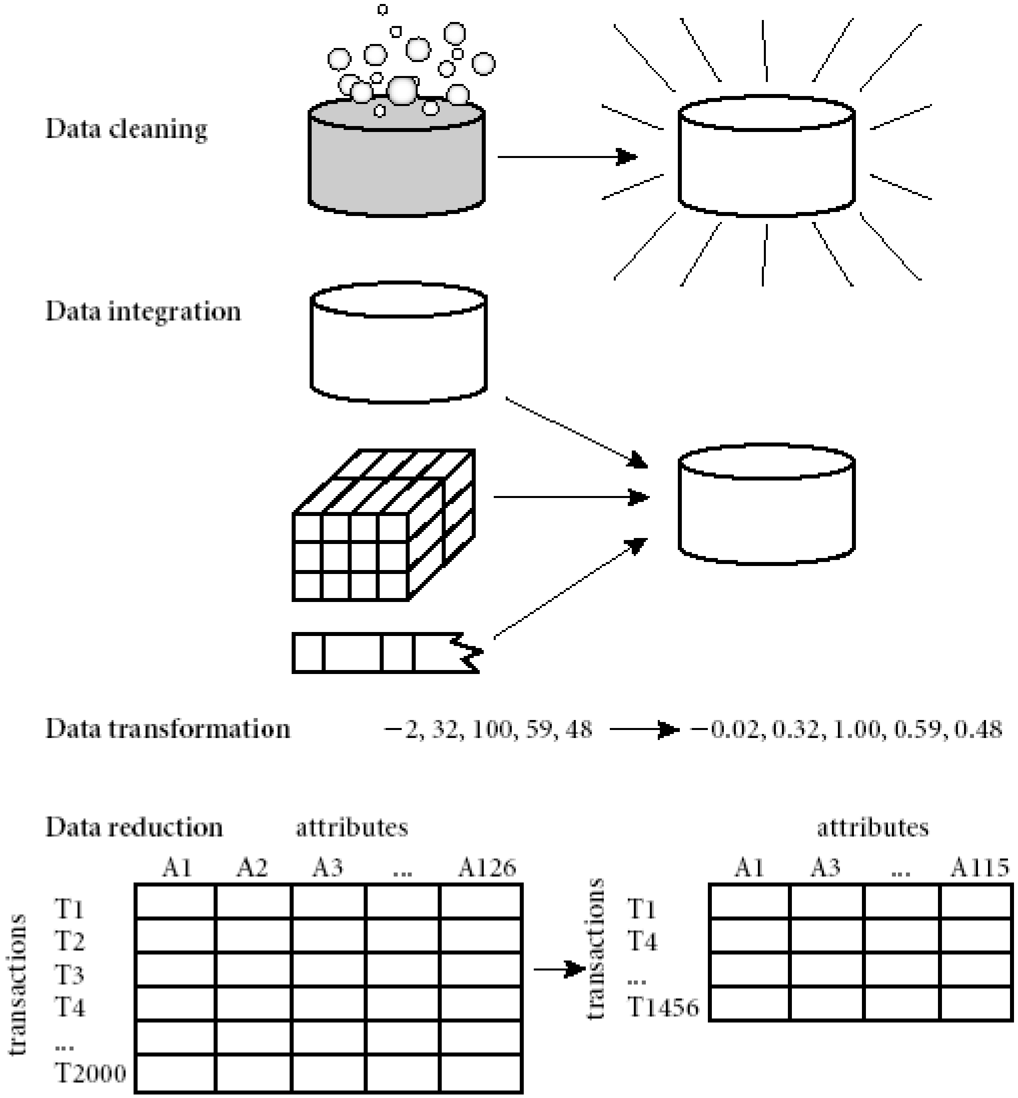
- No quality data, no quality mining results!
- Quality decisions must be based on quality data
  - e.g., duplicate or missing data may cause incorrect or even misleading statistics.

Major Tasks in Data Preprocessing

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 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

Forms of Data Preprocessing



Data cleaning

Data integration

Data transformation −2, 32, 100, 59, 48 → −0.02, 0.32, 1.00, 0.59, 0.48

Data reduction

attributes

transactions

	A1	A2	A3	...	A126
T1					
T2					
T3					
T4					
...					
T2000					

→

attributes

transactions

	A1	A3	...	A115
T1				
T4				
...				
T1456				

References

References

- J. Han, M. Kamber, **Data Mining: Concepts and Techniques**, Elsevier Inc. (2006). (Chapter 2)

The end