Data Preprocessing

——Data Integration and Transformation——

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

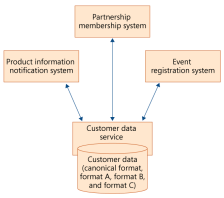
Data Preprocessing

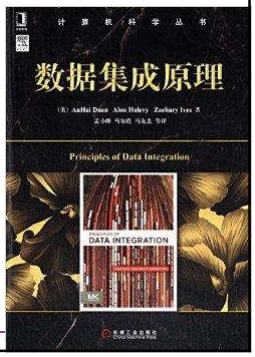
- About data
- Why preprocess the data?
- Descriptive data summarization
- Data cleaning
- Data integration and transformation
- Data reduction
- Discretization and concept hierarchy generation
- Summary

2

Data Integration

- Data integration(数据集成):
  - Combines data from multiple sources into a coherent store
- Schema integration (schema集成)
  - Integrate raw data from different sources
  - Entity identification problem: identify real world entities from multiple data sources, e.g., A.cust-id ≡ B.cust-#
- Detecting and resolving data value conflicts (值冲突)
  - For the same real world entity, attribute values from different sources are different
  - Possible reasons: different representations, different scales, e.g., metric vs. British units

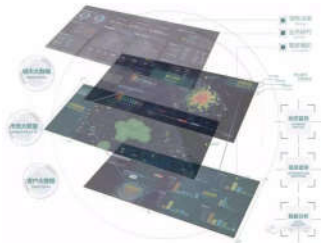


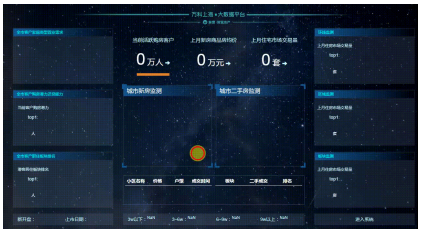


3

Handling Redundancy in Data Integration

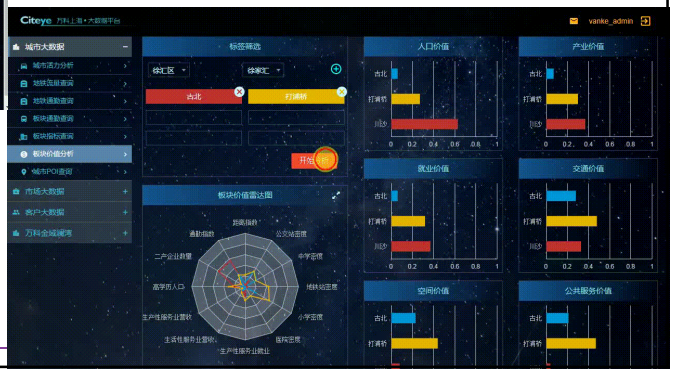
- Redundant data occur often when integration of multiple databases
  - *Object identification*: The same attribute or object may have different names in different databases
  - *Derivable data (导出性数据):* One attribute may be a "derived" attribute in another table, e.g., annual revenue (年度税收)
- Redundant attributes may be able to be detected by *correlation analysis*
- Careful integration of the data from multiple sources may help reduce/avoid redundancies and inconsistencies and improve mining speed and quality





4

Handling Redundancy in Data Integration



5

Correlation Analysis (Numerical Data)

- **Correlation coefficient (also called Pearson's product moment coefficient) (积差相关系数)**

$$r_{A,B} = \frac{\sum (A - \bar{A})(B - \bar{B})}{(n-1)\sigma_A \sigma_B} = \frac{\sum (AB) - n\bar{A}\bar{B}}{(n-1)\sigma_A \sigma_B}$$

where n is the number of tuples, $\bar{A}$ and $\bar{B}$ are the respective means of A and B, σ_A and σ_B are the respective standard deviation of A and B, and Σ(AB) is the sum of the AB cross-product.

- **If $r_{A,B} > 0$, A and B are positively correlated (A's values increase as B's. The higher, the stronger correlation.)**
- **$r_{A,B} = 0$: independent; $r_{A,B} < 0$: negatively correlated**

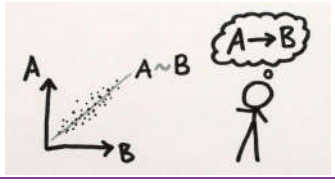
6

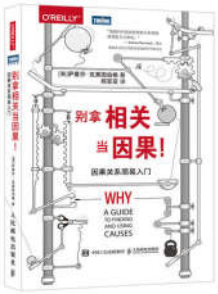
Correlation Analysis (Categorical Data)

- X^2 (chi-square，卡方检验) test

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

- The larger the X^2 value, the more likely the variables are related
- The cells that contribute the most to the X^2 value are those whose actual count is very different from the expected count
- Correlation does not imply causality（因果关系）
  - # of hospitals and # of car-theft in a city are correlated
  - Both are causally linked to the third variable: population（人口）





7

Chi-Square Calculation: An Example

	Play chess	Not play chess	Sum (row)
Like science fiction	250(90)	200(360)	450
Not like science fiction	50(210)	1000(840)	1050
Sum(col.)	300	1200	1500

- X2 (chi-square) calculation (numbers in parenthesis are expected counts calculated based on the data distribution in the two categories)

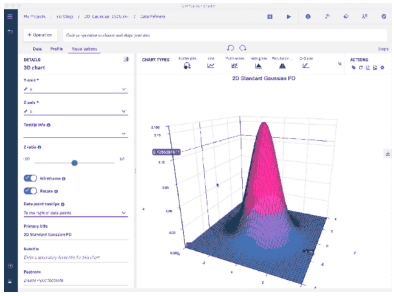
$$\chi^2 = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840} = 507.93$$

- It shows that like_science_fiction（科幻小说） and play_chess are correlated in the group

8

Data Transformation

- Smoothing (平滑) : remove noise from data
- Aggregation (聚合) : summarization, data cube construction
- Generalization (泛化) : concept hierarchy climbing
- Normalization (规范化) : scaled to fall within a small, specified range
  - min-max normalization
  - z-score normalization
  - normalization by decimal scaling
- Attribute/feature construction
  - New attributes constructed from the given ones



9

Data Transformation: Normalization

- min-max normalization

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

- z-score normalization (μ: mean, σ: standard deviation)

$$v' = \frac{v - \mu_A}{\sigma_A}$$

- normalization by decimal scaling

$$v' = \frac{v}{10^j}$$,where j is the smallest integer such that Max(|v'|) < 1

10

Thanks !