Data Mining

2.4 Data Integration

Fall 2008

Instructor: Dr. Masoud Yaghini

Data Integration

- Data integration:
  - Combines data from multiple databases into a coherent store
  - Denormalization tables (often done to improve performance by avoiding joins)
- Integration of the data from multiple sources may produces redundancies and inconsistencies in the resulting data set.
- Tasks of data integration:
  - Detecting and resolving data value and schema conflicts
  - Handling Redundancy in Data Integration

Outline

- Detecting and Resolving Data Value and Schema Conflicts
- Handling Redundancy in Data Integration
- References

Detecting and Resolving Data Value and Schema Conflicts

Schema Integration

- Schema Integration:
  - Integrate metadata from different sources
  - The same attribute or object may have different names in different databases
  - e.g. *customer_id* in one database and *cust_number* in another
- The metadata include:
  - the name, meaning, data type, and range of values permitted for the attribute, and etc.

Detecting and resolving data value conflicts

- For the same real world entity, attribute values from different sources are different
- This may be due to differences in representation, scaling, or encoding.
- Examples:
  - the data codes for ***pay_type*** in one database may be *"H"* and *"S"*, and *1* and *2* in another.
  - a ***weight*** attribute may be stored in metric units in one system and British imperial units in another.
  - For a hotel chain, the ***price*** of rooms in different cities may involve not only different currencies but also different services (such as free breakfast) and taxes.

Detecting and resolving data value conflicts

- This step also relates to data cleaning, as described earlier.

Handling Redundancy in Data Integration

Handling Redundancy in Data Integration

- Redundant data occur often when integration of multiple databases
- One attribute may be a “derived” attribute in another table,
  - e.g., Age=“19” and Birth_year =“1990”
- Redundant attributes may be able to be detected by correlation analysis

Correlation Analysis (Numerical Data)

- Correlation coefficient (also called Pearson's product moment coefficient)

$$r_{A,B} = \frac{\sum_{i=1}^{N}(a_i - \overline{A})(b_i - \overline{B})}{N\sigma_A\sigma_B} = \frac{\sum_{i=1}^{N}(a_i b_i) - N\overline{A}\,\overline{B}}{N\sigma_A\sigma_B}$$

- where
  - n is the number of tuples
  - $\overline{A}$ and $\overline{B}$ are the respective means of A and B
  - σ_A and σ_B are the respective standard deviation of A and B
  - $\Sigma(a_i b_i)$ is the sum of the AB cross-product

Correlation Analysis (Numerical Data)

- 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

- Correlation does not imply causality
  - # of hospitals and # of car-theft in a city are correlated
  - Both are causally linked to the third variable: population

Correlation Analysis (Categorical Data)

- A correlation relationship between two categorical (discrete) attributes, *A* and *B*, can be discovered by a X^2 (**chi-square**) test.

Correlation Analysis (Categorical Data)

- Suppose:
  - *A* has *c* distinct values, namely *a1, a2, …, ac.*
  - *B* has *r* distinct values, namely *b1, b2, …, br*
  - The data tuples described by *A* and *B* can be shown as a contingency table, with
    - the *c* values of *A* making up the columns and
    - the *r* values of *B* making up the rows.
  - Let *(Ai, Bj)* denote the event that attribute *A* takes on value *ai* and attribute *B takes on value bj, that is, where* $(A = a_i, B = b_j)$.
  - Each and every possible (A_i, B_j) joint event has its own cell (or slot) in the table.

Correlation Analysis (Categorical Data)

- The X^2 value (also known as the *Pearson* X^2 statistic) is computed as:

$$\chi^2 = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

  – where o_{ij} is the observed frequency (i.e., actual count) of the joint event *(A_i, B_j) and*

  – e_{ij} is the expected frequency of *(A_i, B_j)*, which can be computed as

$$e_{ij} = \frac{count(A = a_i) \times count(B = b_j)}{N}$$

Correlation Analysis (Categorical Data)

- where
  - N is the number of data instances, $count(A{=}a_i)$ is the number of tuples having value a_i *for A*
  - $count(B = b_j)$ is the number of tuples having value b_j for B.

- 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

Chi-Square Calculation: An Example

- Suppose that a group of 1,500 people was surveyed.
- The observed frequency (or count) of each possible joint event is summarized in the contingency table shown in the Table

	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

- The numbers in parentheses are the expected frequencies (calculated based on the data distribution for both attributes using Equation e_{ij}).
- Are ***like_science_fiction*** and ***play_chess*** correlated?

Chi-Square Calculation: An Example

- For example, the expected frequency for the cell (play_chess, fiction) is

$$e_{11} = \frac{count(play\_chess) * count(like\_science\_fiction)}{N} = \frac{300 * 450}{1500} = 90$$

- Notice that
  - the sum of the expected frequencies must equal the total observed frequency for that row, and
  - the sum of the expected frequencies in any column must also equal the total observed frequency for that column.

Chi-Square Calculation: An Example

- We can get X^2 by:

$$
\begin{aligned}
\chi^2 &= \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840} \\
&= 284.44 + 121.90 + 71.11 + 30.48 = 507.93.
\end{aligned}
$$

- For this 2 x 2 table, the degrees of freedom are (2-1)(2-1) = 1.
- For 1 degree of freedom, the X^2 value needed to reject the hypothesis at the 0.001 significance level is 10.828 (taken from the table of upper percentage points of the X^2 distribution, typically available from any textbook on statistics).

Correlation Analysis (Categorical Data)

- Since our computed value is above this, we can reject the hypothesis that ***play chess*** and ***preferred reading*** are independent and conclude that the two attributes are (strongly) correlated for the given group of people.

References

References

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

The end