Data Integration and Data Transformation

Combines the data from multiple data sources- Data Integration

What are the data sources?

Problems in the data integration

1. Entity Identification Problem
2. Redundancy

Entity Identification Problem

Customer Id – Database1
Customer Number – Database2

Merging the values from database1 and database 2 is difficult due to different names of attributes in the customer entity.

Customer table with customer id
Customer_Details table with customer id

Solution of Entity Identification Problem - Metadata

Redundnacy Problem

Customer Number – Database1
Customer Number – Database2

Merging the values from database1 and database 2 is difficult due to different names of attributes in the customer entity.

Customer table with customer id
Customer_Details table with customer id

Solution of Redundancy Problem - Correlation

Correlation

Pearson's Chi-Square Test

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

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

Correlation : Example Problem

Pearson's Chi-Square Test

	male	*female*	Total
fiction	250 (90)	200 (360)	450
non_fiction	50 (210)	1000 (840)	1050
Total	300	1200	1500

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

Data Transformation

Smoothing
- 1. Binning
- 2. Regression
- 3. Clustering

Aggregation
- Summary or Aggregation operations are applied to the data

Generalization
- Use CH

Normalization
- Attribute Data are scaled
  - 1. Min-Max Normalization
  - 2. Z-score Normalization
  - 3. Decimal Scaling

Attribute Construction

Min-Max Normalization and Example

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

Min-max normalization. Suppose that the minimum and maximum values for the attribute *income* are \$12,000 and \$98,000, respectively. We would like to map *income* to the range $[0.0, 1.0]$. By min-max normalization, a value of \$73,600 for *income* is transformed to $\frac{73{,}600-12{,}000}{98{,}000-12{,}000}(1.0-0)+0=0.716$. ■

Z-Score Normalization and Example

$$v' = \frac{v - \bar{A}}{\sigma_A}$$

z-score normalization Suppose that the mean and standard deviation of the values for the attribute *income* are \$54,000 and \$16,000, respectively. With z-score normalization, a value of \$73,600 for *income* is transformed to $\frac{73,600-54,000}{16,000} = 1.225$. ■

Decimal Scaling and Example

$$v' = \frac{v}{10^j}$$

Decimal scaling. Suppose that the recorded values of A range from -986 to 917. The maximum absolute value of A is 986. To normalize by decimal scaling, we therefore divide each value by 1,000 (i.e., $j = 3$) so that -986 normalizes to -0.986 and 917 normalizes to 0.917. ■