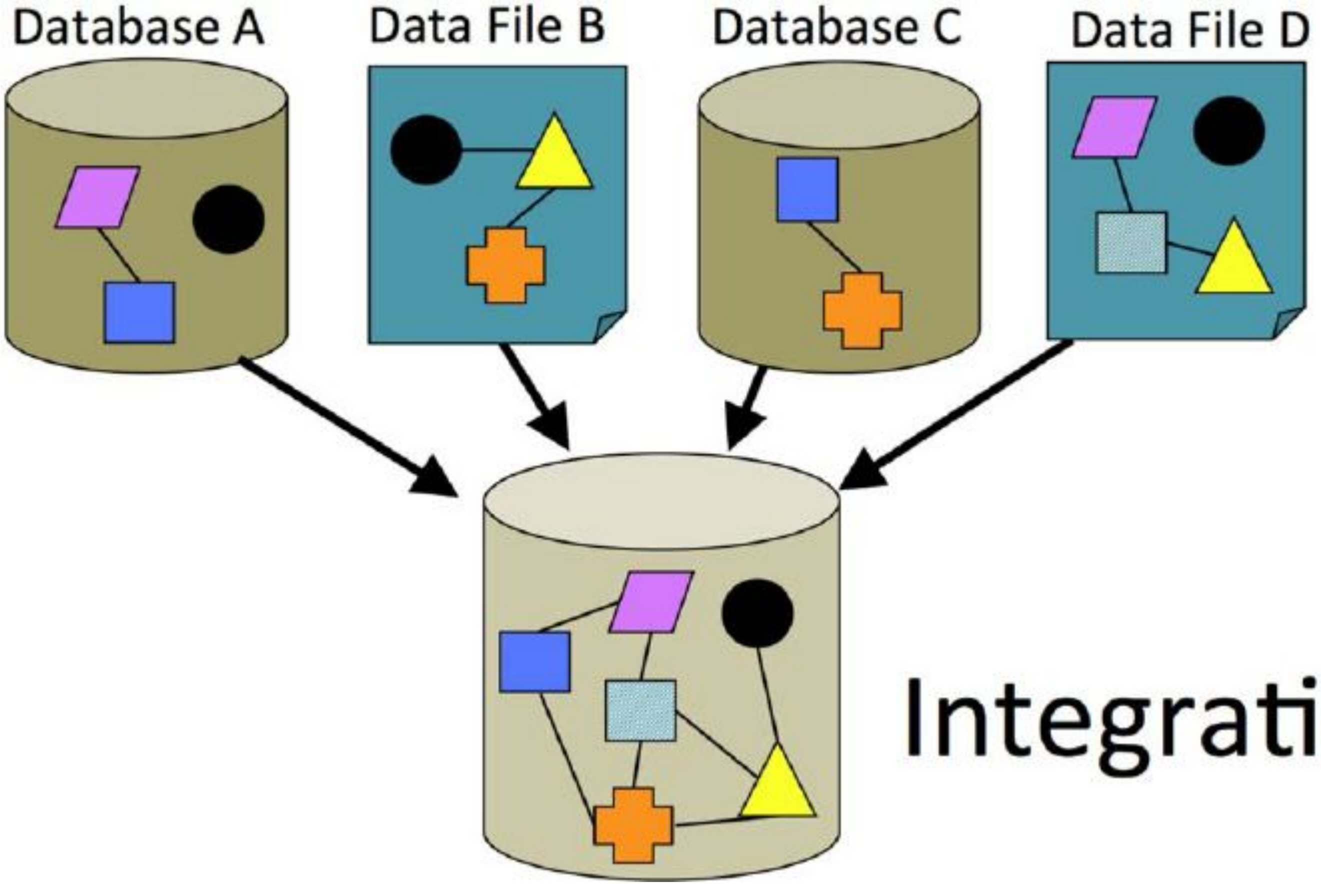
Data Integration

Lesson 3b



Integration

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:* One attribute may be a "derived" attribute in another table, e.g., annual revenue
- 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

Correlation Analysis

Numerical Data :

Pearson’s Correlation Coefficient

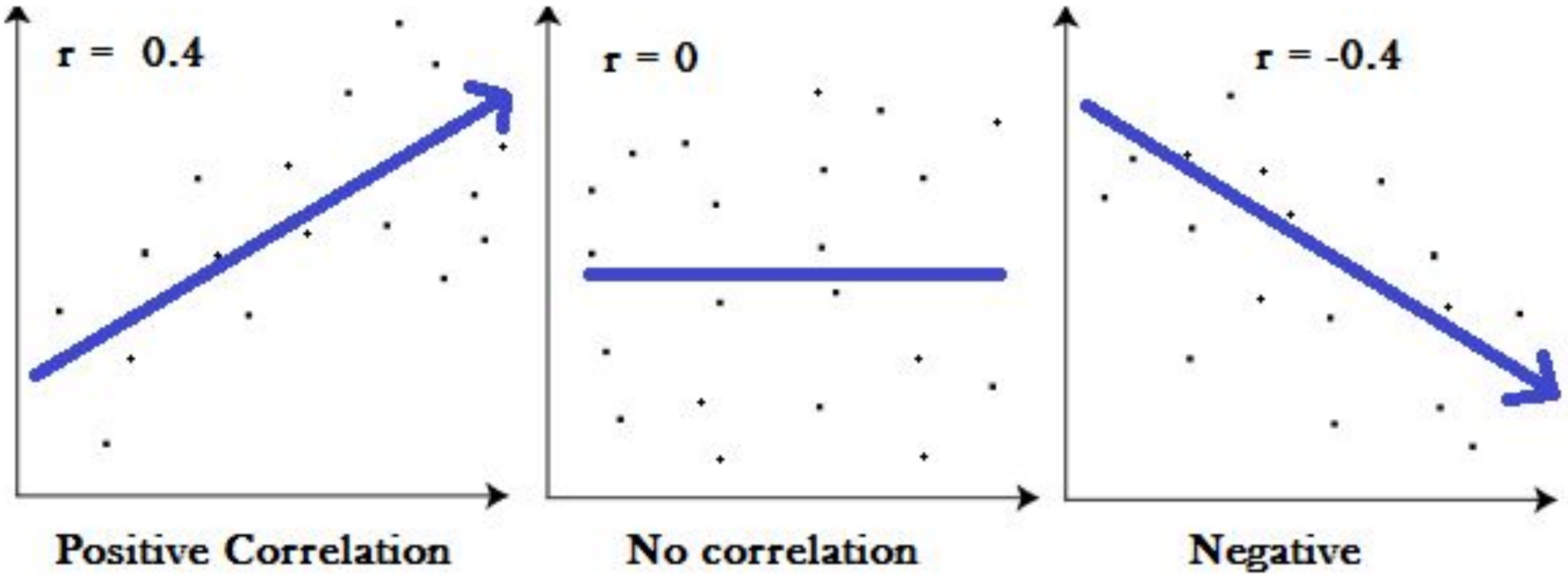
Categorical Data :

Chi Square Test

Correlation analysis : Numerical data
(Pearson’s Correlation Coefficient)

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}}$$

- > 0 , A and B positively correlated
  - values of A increase as values of B increase
  - The higher the value, the more each attribute implies the other
  - High value indicate that A (or B) may be removed as a redundancy
- = 0, A and B independent (no correlation)
- < 0, A and B negatively correlated
  - Values of one attribute increase as the values of the other attribute decrease (discourages each other)


r = 0.4
r = 0
r = -0.4
Positive Correlation
No correlation
Negative

- Find the value of the correlation coefficient from the table:

Subject	Age	Glucose level
1	43	99
2	21	65
3	25	79
4	42	75
5	57	87
6	59	81

Solution

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}}$$

SUBJECT	AGE X	GLUCOSE LEVEL Y	XY	X^2	Y^2
1	43	99	4257	1849	9801
2	21	65	1365	441	4225
3	25	79	1975	625	6241
4	42	75	3150	1764	5625
5	57	87	4959	3249	7569
6	59	81	4779	3481	6561
Σ	247	486	20485	11409	40022

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}}$$

From our table:

- $\Sigma x = 247$
- $\Sigma y = 486$
- $\Sigma xy = 20{,}485$
- $\Sigma x^2 = 11{,}409$
- $\Sigma y^2 = 40{,}022$
- n is the sample size, in our case = 6

The correlation coefficient =

- $6(20{,}485) - (247 \times 486) / [\sqrt{[[6(11{,}409) - (247^2)] \times [6(40{,}022) - 486^2]]}$
  $= 0.5298$

The range of the correlation coefficient is from -1 to 1. Our result is 0.5298 or 52.98%, which means the variables have a moderate positive correlation

Correlation analysis (categorical data)

- X^2 (chi-square) test

$$\chi^2_{n-1} = \sum_{i=1}^{n} \frac{(Observed_i - Expected_i)^2}{Expected_i}$$

- n is the number of possible values
- 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
  - # of hospitals and # of car-theft in a city are correlated
  - Both are causally linked to the third variable: population

Observed Counts

Gender	Drank Alcohol in Last 2 Hours? Yes	No	Total
Male	77	404	481
Female	16	122	138
Total	93	526	619

Expected Counts

Gender	Drank Alcohol in Last 2 Hours? Yes	No	Total
Male	(93*481)/619=72.3	(526*481)/619=408.7	481
Female	(93*138)/619=20.7	(526*138)/619=117.3	138
Total	93	526	619

Example:

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

Probability to play chess: P(chess) = 300/1500 = 0.2

Probability to like science fiction: P(SciFi) = 450/1500 = 0.3

If science fiction and chess playing are independent attributes, then the probability to like SciFi AND play chess is

P(SciFi, chess) = P(SciFi) · P(chess) = 0.06

That means, we expect 0.06 · 1500 = 90 such cases (if they are independent)

	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

- X^2 (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 and play_chess are correlated in the group

Next Topic: Association Rule Mining