Data Integration — 2

Faculty of Information Technology, Monash University, Australia

FIT5196 week 11

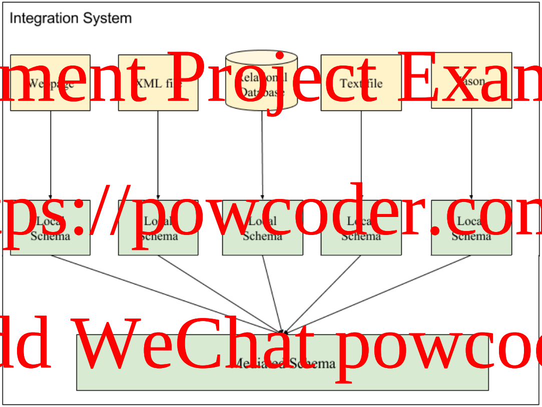
Data Integration

MONASH University

- What is Data Integration?
  - A process in which heterogeneous data is retrieved and combined as an incorporated form and structure.
- What is the goal of Data Integration?
  - Create a single representation that provides a more accurate description than any of the individual data sources

Data Integration: Schema Integration

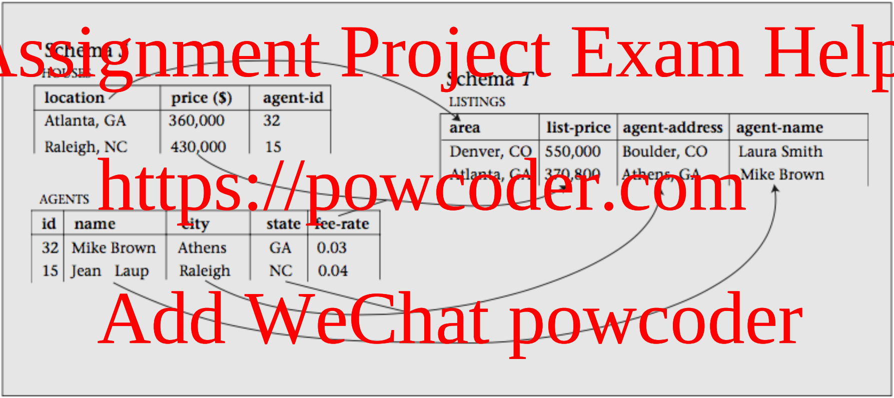
MONASH University



- Data always comes from different sources.
- Each source has its own schemas and references to objects, even though these sources might model the same domain.
- Often, users directly interacts with the mediation schema instead of local

Schema Integration: Structure & Name Conflicts

MONASH University



Schema S

HOUSES

location	price ($)	agent-id
Atlanta, GA	360,000	32
Raleigh, NC	430,000	15

AGENTS

id	name	city	state	fee-rate
32	Mike Brown	Athens	GA	0.03
15	Jean Laup	Raleigh	NC	0.04

Schema T

LISTINGS

area	list-price	agent-address	agent-name
Denver, CO	550,000	Boulder, CO	Laura Smith
Atlanta, GA	370,800	Athens, GA	Mike Brown

Figure 2. The Schemas of Two Relational Databases S and T on House Listing, and the Semantic Correspondences between Them.

Figure is from "Semantic-Integration Research in the Database community" by AnHai Doan and Alon Y. Halevy

Schema Integration: Semantic Matching

DVD-VENDOR

Movies(id, title, year)

Products(mid, releaseDate, releaseCompany, basePrice, rating, saleLocID)

Locations(lid, name, taxRate)

AGGREGATOR

Items(name, releaseInfo, classification, price)

FIGURE 5.1 Example of two database schemas. Schema DVD-VENDOR belongs to a DVD vendor, while AGGREGATOR belongs to a shopping site that aggregates products from multiple vendors.

Figure is from chapter 5 of "Principles of data integration"

- One-to-One match
  - Movies.title $\approx$ Items.name
  - Movies.year $\approx$ Items.year
  - Product.rating $\approx$ Items.classification
- One-to-Many match
  - Items.price $\approx$ Products.basePrices $\times$ (1 + Locations.taxRate)

Outline

MONASH University

1 Recap

2 Data-Level Integration
- Attribute-Level Integration
- Tuple-Level Integration

3 Summary

Data-Level Integration

- Data-Level Integration: related to the integrated contents/values of data not the schema
- Categories
  - Attribute-level (columns)
    - Redundancy
    - Correlation
  - Tuple-level (rows)
    - Duplication
    - Inconsistency

Data-Level Integration: Attribute-Level Issues

MONASH University

- Problems: combining different data sources might result in a redundant representation
- Examples
  - When any of the attributes can be calculated from others
    - e.g., annual salary from fortnight payment
  - When different values represent the same attribute but with different units
    - e.g., weight in kg and lb
- Techniques to find correlation between attributes
  - Chi-square Test for categorial varaibles
  - Correlation Coefficient for numerical attributes

Attribute-Level Issues: Chi-Sqaure Test

- Chi-square test for categorial variables
  - Test for independence compares two variables in a contingency table to see if they are related.
  - Hypothesis statements:
    - Null Hypothesis: The two categorical variables are independent.
    - Alternative Hypothesis: The two categorical variables are dependent.
  - The chi-square test statistic

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

where

- O represents the observed frequency.
- E is the expected frequency under the null hypothesis:

$$E = \frac{row\_total \times column\_total}{sample\_size}$$

Attribute-Level Issues: Chi-Sqaure Test

MONASH University

- Chi-square test for categorial variables: Is gender independent of education level?

	High School	Bachelors	Masters	Ph.d.	Total
Female	60	54	46	41	201
Male	40	44	53	57	194
Total	100	98	99	98	395

	High School	Bachelors	Masters	Ph.d.	Total
Female	50.886	49.868	50.377	49.868	201
Male	49.114	48.132	48.623	48.132	194
Total	100	98	99	98	395

Attribute-Level Issues: Chi-Sqaure Test

MONASH University

- Chi-square test for categorial variables: Is gender independent of education level?

	High School	Bachelors	Masters	Ph.d.	Total
Female	60	54	46	41	201
Male	40	44	53	57	194
Total	100	98	99	98	395

	High School	Bachelors	Masters	Ph.d.	Total
Female	50.886	49.868	50.377	49.868	201
Male	49.114	48.132	48.623	48.132	194
Total	100	98	99	98	395

- Null Hypothesis: Gender and Education Level are independent.
- Alternative Hypothesis: Gender and Education Level are dependent

Attribute-Level Issues: Chi-Sqaure Test

MONASH University

- Chi-square test for categorial variables: Is gender independent of education level?

	High School	Bachelors	Masters	Ph.d.	Total
Female	60	54	46	41	201
Male	40	44	53	57	194
Total	100	98	99	98	395

	High School	Bachelors	Masters	Ph.d.	Total
Female	50.886	49.868	50.377	49.868	201
Male	49.114	48.132	48.623	48.132	194
Total	100	98	99	98	395

$$50.886 = \frac{100 \times 201}{395}$$

Attribute-Level Issues: Chi-Sqaure Test

MONASH University

- Chi-square test for categorial variables: Is gender independent of education level?

	High School	Bachelors	Masters	Ph.d.	Total
Female	60	54	46	41	201
Male	40	44	53	57	194
Total	100	98	99	98	395

	High School	Bachelors	Masters	Ph.d.	Total
Female	50.886	49.868	50.377	49.868	201
Male	49.114	48.132	48.623	48.132	194
Total	100	98	99	98	395

$$\chi^2 = \frac{(60 - 50.886)^2}{50.886} + \frac{(54 - 49.868)^2}{49.868} + \cdots = 8.006$$

Attribute-Level Issues: Chi-Sqaure Test

- Chi-square test for categorial variables: Is gender independent of education level?

	High School	Bachelors	Masters	Ph.d	Total
Female	60	54	46	41	201
Male	40	44	53	57	194
Total	100	98	99	98	395

	High School	Bachelors	Masters	Ph.d.	Total
Female	50.886	49.868	50.377	49.868	201
Male	49.114	48.132	48.623	48.132	194
Total	100	98	99	98	395

- $\chi^2 = 8.006 > 7.815$ (The critical value of χ^2 with 3 degree of freedom)
- Reject the null hypothesis and conclude that the education level depends on gender at a 5% level of significance

Attribute-Level Issues: Chi-Sqaure Test

- Chi-square test for categorial variables: Is gender independent of education level?

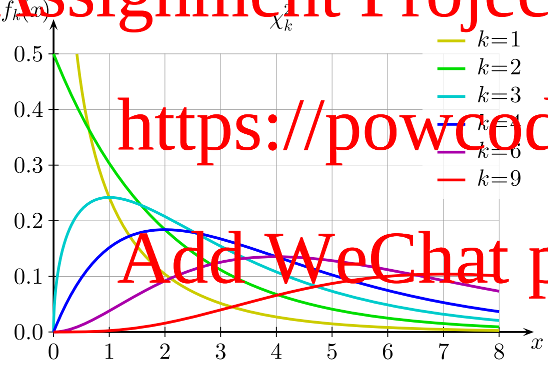
Percentage Points of the Chi-Square Distribution

Degrees of Freedom	Probability of a larger value of x ²								
	0.99	0.95	0.90	0.75	0.50	0.25	0.10	0.05	0.01
1	0.000	0.004	0.016	0.102	0.455	1.32	2.71	3.84	6.63
2	0.020	0.103	0.211	0.575	1.386	2.77	4.61	5.99	9.21
3	0.115	0.352	0.584	1.212	2.366	4.11	6.25	7.81	11.34
4	0.297	0.711	1.064	1.923	3.357	5.39	7.78	9.49	13.28
5	0.554	[illegible]	[illegible]	[illegible]	4.351	[illegible]	9.24	[illegible]	15.09
6	0.872	[illegible]	[illegible]	[illegible]	5.348	[illegible]	[illegible]	[illegible]	[illegible]
7	1.239	[illegible]	[illegible]	[illegible]	6.346	9.04	12.02	[illegible]	18.48
8	1.647	2.733	3.490	5.071	7.344	10.22	13.36	15.51	20.09
9	2.088	3.325	4.168	5.899	8.343	11.39	14.68	16.92	21.67
10	2.558	3.940	4.865	6.737	9.342	12.55	15.99	18.31	23.21
11	3.053	4.575	5.578	7.584	10.341	13.70	17.28	19.68	24.72
12	3.571	5.226	6.304	8.438	11.340	14.85	18.55	21.03	26.22
13	4.107	5.892	7.042	[illegible]	12.340	15.98	[illegible]	22.36	27.69
14	4.660	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	29.14
15	5.229	[illegible]	[illegible]	11.037	[illegible]	[illegible]	[illegible]	[illegible]	30.58
16	5.812	7.962	9.312	11.912	15.338	19.37	23.54	26.30	32.00
17	6.408	8.672	10.085	12.792	16.338	20.49	24.77	27.59	33.41
18	7.015	9.390	10.865	13.675	17.338	21.60	25.99	28.87	34.80
19	7.633	10.117	11.651	14.562	18.338	22.72	27.20	30.14	36.19
20	8.260	10.851	12.443	15.452	19.337	23.83	28.41	31.41	37.57
22	9.542	12.338	14.041	17.240	21.337	26.04	30.81	33.92	40.29
24	10.856	13.848	15.659	19.037	23.337	28.24	33.20	36.42	42.98
26	12.198	15.379	17.292	20.843	25.336	30.43	35.56	38.89	45.64
28	13.565	16.928	18.939	22.657	27.336	32.62	37.92	41.34	48.28
30	14.953	18.493	20.599	24.478	29.336	34.80	40.26	43.77	50.89
40	22.164	26.509	29.051	33.660	39.335	45.62	51.80	55.76	63.69
50	27.707	34.764	37.689	42.942	49.335	56.33	63.17	67.50	76.15
60	37.485	43.188	46.459	52.294	59.335	66.98	74.40	79.08	88.38

- $\chi^2 = 8.006$
- The degree of freedom: $(r-1)(c-1) = 3$
- The critical value of χ^2 at a 5% level of significance : 7.815

Attribute-Level Issues: Chi-Sqaure Test

- Chi-square test for categorial variables: Is gender independent of education level?



- $\chi^2 = 8.006 > 7.815$
- Reject the null hypothesis and conclude that the education level depends on gender at a 5% level of significance

Attribute-Level Issues: Correlation Coefficient

- Correlation Coefficient, r, also called Pearson correlation coefficient
  - Measures the strength and the direction of a linear relationship between two variables
  - Compute r

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

  - r values:
    - The value of r is such that $-1 < r < +1$
    - Positive correlation: If x and y have a strong positive linear correlation, r is close to +1.
    - Negative correlation: If x and y have a strong negative linear correlation, r is close to -1.
    - No correlation: If there is no linear correlation or a weak linear correlation, r is close to 0.

Attribute-Level Issues: Coefficient of determination

MONASH University

- Coefficient of determination
  - The proportion of the variance (fluctuation) of one variable that is predictable from the other variable.
  - $0 < r^2 < 1$ denotes the strength of the linear association between x and y.
  - The coefficient of determination is a measure of how well the regression line represents the data. If the regression line passes exactly through every point on the scatter plot, it would be able to explain all of the variation. The further the line is away from the points, the less it is able to explain.

Attribute-Level Issues: Coefficient of determination

MONASH University

	x	y	xy	x^2	y^2
	313000	1340	419420000	97969000000	1795600
	2384000	3650	8701600000	5.68346E+12	13322500
	342000	1930	660060000	1.16964E+11	3724900
	420000	2000	840000000	1.764E+11	4000000
	550000	1940	1067000000	3.025E+11	3763600
	490000	880	431200000	2.401E+11	774400
	335000	1350	452250000	1.12225E+11	1822500
	482000	2710	1306220000	2.32324E+11	7344100
	452500	2430	1099575000	2.04756E+11	5904900
	640000	1520	972800000	4.096E+11	2310400
	463000	1710	791730000	2.14369E+11	2924100
	1400000	2920	4088000000	1.96E+12	8526400
	588500	2330	1371205000	3.46332E+11	5428900
	365000	1090	397850000	1.33225E+11	1188100
	1200000	2910	3492000000	1.44E+12	8468100
	242500	1200	291000000	58806250000	1440000
	419000	1570	657830000	1.75561E+11	2464900
	285000	2200	627000000	81225000000	4840000
	367500	3110	1142925000	1.35056E+11	9672100
Sum	11739000	38790	28809665000	1.21209E+13	89715500

$$r = \frac{n\sum(xy) - (\sum x)(\sum y)}{\sqrt{n(\sum x^2) - (\sum x)^2}\sqrt{n(\sum y^2) - (\sum y)^2}} = 0.676747624$$

Attribute-Level Issues: Coefficient of determination

	x	y	xy	x^2	y^2
	313000	1340	419420000	97969000000	1795600
	2384000	3650	8701600000	5.68346E+12	13322500
	342000	1930	660060000	1.16964E+11	3724900
	420000	2000	840000000	1.764E+11	4000000
	550000	1940	1067000000	3.025E+11	3763600
	490000	880	431200000	2.401E+11	774400
	335000	1350	452250000	1.12225E+11	1822500
	482000	2710	1306220000	2.32324E+11	7344100
	452500	2430	1099575000	2.04756E+11	5904900
	640000	1520	972800000	4.096E+11	2310400
	463000	1710	791730000	2.14369E+11	2924100
	1400000	2920	4088000000	1.96E+12	8526400
	588500	2330	1371205000	3.46332E+11	5428900
	365000	1090	397850000	1.33225E+11	1188100
	1200000	2910	3492000000	1.44E+12	8468100
	242500	1200	291000000	58806250000	1440000
	419000	1570	657830000	1.75561E+11	2464900
	285000	2200	627000000	81225000000	4840000
	367500	3110	1142925000	1.35056E+11	9672100
Sum	11739000	38790	28809665000	1.21209E+13	89715500

$$r^2 = 0.676747624^2 = 0.457987347$$

Attribute-Level Issues: Coefficient of determination

MONASH University

	x	y	xy	x^2	y^2
	313000	1340	419420000	97969000000	1795600
	2384000	3650	8701600000	5.68346E+12	13322500
	342000	1930	660060000	1.16964E+11	3724900
	420000	2000	840000000	1.764E+11	4000000
	550000	1940	1067000000	3.025E+11	3763600
	490000	880	431200000	2.401E+11	774400
	335000	1350	452250000	1.12225E+11	1822500
	482000	2710	1306220000	2.32324E+11	7344100
	452500	2430	1099575000	2.04756E+11	5904900
	640000	1520	972800000	4.096E+11	2310400
	463000	1710	791730000	2.14369E+11	2924100
	1400000	2920	4088000000	1.96E+12	8526400
	588500	2330	1371205000	3.46332E+11	5428900
	365000	1090	397850000	1.33225E+11	1188100
	1200000	2910	3492000000	1.44E+12	8468100
	242500	1200	291000000	58806250000	1440000
	419000	1570	657830000	1.75561E+11	2464900
	285000	2200	627000000	81225000000	4840000
	367500	3110	1142925000	1.35056E+11	9672100
Sum	11739000	38790	28809665000	1.21209E+13	89715500

- Correlation vs Causality

Attribute-Level Issues: Coefficient of determination

MONASH University

- Regression Sum of Squares (SSR) (or explained sum of squares)

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

- Residual Sum of squares (RSS)

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2$$

- Total sum of squares (TSS)

$$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

- R^2 is defined as

$$R^2 = 1 - \frac{RSS}{TSS}$$

Attribute-Level Issues: Coefficient of determination

MONASH University

- Regression Sum of Squares (SSR) (or explained sum of squares)

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

- Residual Sum of squares (RSS)

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2$$

- Total sum of squares (TSS)

$$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

- Question:

$$TSS \stackrel{?}{=} SSR + RSS$$

Attribute-Level Issues: Coefficient of determination

MONASH University

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 \stackrel{?}{=} \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad (1)$$

$$\begin{aligned}\sum_{i=1}^{n}(y_i - \bar{y})^2 &= \sum_{i=1}^{n}(y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2 \\ &= \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + 2\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) \qquad (2)\end{aligned}$$

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) \stackrel{?}{=} 0 \qquad (3)$$

Data-Level Integration: Tuple-Level Integration

- Duplicates
  - Two or more rows (i.e., tuples) refer to the same object.
- Inconsistent update
  - Duplicated records are not updated simultaneously.
- Issues with tuple-level integration
  - Formatting convertors
  - Different naming conventions
  - ⋮

- Tuple Matching methods
  - String Matching
  - Data Matching

Tuple-Level Integration: String Matching

MONASH University

- Problems: Given two sets of strings X and Y, find all pairs of strings (x, y), where $x \in X$ and $y \in Y$, such that x and y refer to the same entity.

Set X
x_1=Dave Smith
x_2=Joe Wilson
x_3=Dan Smith

(a)

Set Y
y_1=David D. Smith
y_2=Daniel W. Smith

(b)

Matches
(x_1, y_1)
(x_3, y_2)

(c)

Figure is from chapter 4 of "Principles of Data Integration"

Tuple-Level Integration: String Matching

- Methods: Similarity Measures
  - Sequence-based Similarity Measures: View strings as sequences of characters, compute a cost of transforming one string into the other.
    - Edit Distance
    - The Needleman-Wunch measure
    - The Affine Gap measure
    - The Smith-Waterman measure
    - ⋮
  - Set-based Similarity Measures: View strings as sets or multi-sets of tokens, and use set-related properties to compute similarity scores.
    - The Overlap measure
    - The TF/IDF measure
  - Hybrid Similarity Measures: combines sequence-based and set-based measures
    - The Generalised Jaccard measure
    - The Soft TF/IDF measure
  - Phonetic Similarity Measure: matches strings based on their sound.

String Matching: Edit Distance

- The minimum edit distance between two strings
- Is minimum number of editing operations
  - Insertion
  - Deletion
  - Substitution
- Needed to transform one to another

String Matching: Edit Distance

$$d(i,j) = \min \begin{cases} d(i-1, j-1) & \text{if } x_i = y_j \quad \text{// copy} \\ d(i-1, j-1) + 1 & \text{if } x_i <> y_j \quad \text{// substitute} \\ d(i-1, j) + 1 & \text{// delete } x_i \\ d(i, j-1) + 1 & \text{// insert } y_j \end{cases}$$

(a)

$$d(i,j) = \min \begin{cases} d(i-1, j-1) + c(x_i, y_j) & \text{// copy or substitute} \\ d(i-1, j) + 1 & \text{// delete } x_i \\ d(i, j-1) + 1 & \text{// insert } y_j \end{cases}$$

$$c(x_i, y_j) = \begin{cases} 0 & \text{if } x_i = y_j \\ 1 & \text{otherwise} \end{cases}$$

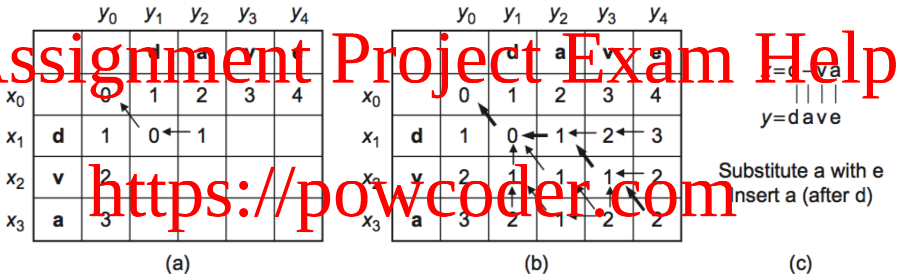
(b)

Figure [illegible] from chapter 4 of “Principles of Data Integration”

Transform string $x_1, \ldots, x_i, \ldots, x_n$ to $y_1, \ldots, y_j, \ldots, y_m$

- Transform $x_1, \ldots, x_{i-1}$ into $y_1, \ldots, y_{j-1}$ if $x_i = y_j$
- Transform $x_1, \ldots, x_{i-1}$ into $y_1, \ldots, y_{j-1}$, then substituting x_i with y_i if $x_i \neq y_j$
- Deleting x_i, then transform $x_1, \ldots, x_{i-1}$ into $y_1, \ldots, y_j$,
- Transform $x_1, \ldots, x_i$ into $y_1, \ldots, y_{j-1}$, then insert y_j

String Matching: Edit Distance



$$
d(i,j) = \min \begin{cases} d(i-1, j-1) + c(x_i, y_j) & \text{// copy or substitute} \\ d(i-1, j) + 1 & \text{// delete } x_i \\ d(i, j-1) + 1 & \text{// insert } y_j \end{cases}
$$

$$
c(x_i, y_j) = \begin{cases} 0 & \text{if } x_i = y_j \\ 1 & \text{otherwise} \end{cases}
$$

Figure is from chapter 4 of "Principles of Data Integration"

String Matching: The Needleman-Wunch Measure


```
d--va
|  ||
deeve
```

(a)

	d	a	v	e
d	2	−1	−1	−1
v	−1	2	−1	−1
a	−1	−1	2	−1
e	−1	−1	−1	2

$c_g = 1$

(b)

Figure is from chapter 4 of "Principles of Data Integration"

String Matching: The Needleman-Wunch Measure

MONASH University

$$s(i,j) = \max \begin{cases} s(i-1, j-1) + c(x_i, y_j) \\ s(i-1, j) - c_g \\ s(i, j-1) - c_g \end{cases}$$

$$s(0, j) = -jc_g$$
$$s(i, 0) = -ic_g$$

(a)

		d	e	e	v	e
	0	–1	–2	–3	–4	–5
d	–1	2	1	0	–1	–2
v	–2	1	1	0	2	[illegible]
a	–3	0	0	0	1	1

(b)

```
d--va
|  ||
deeve
```

(c)

Figure is from chapter 4 of "Principles of Data Integration"

Tuple-Level Integration: The TF/IDF measures

MONASH University

$x = aab \Rightarrow B_x = \{a, a, b\}$

$y = ac \Rightarrow B_y = \{a, c\}$

$z = a \Rightarrow B_z = \{a\}$

(a)

$tf(a, x) = 2$

$tf(b, x) = 1$

...

$tf(c, z) = 0$

$idf(a) = 3/3 = 1$

$idf(b) = 3/1 = 3$

$idf(c) = 3/1 = 3$

(b)

	a	b	c
v_x	2	3	0
v_y	3	0	3
v_z	3	0	0

(c)

Figure is from chapter 4 of "Principles of Data Integration"

$$s(p, q) = \frac{\sum_{t \in T} v_p(t) \cdot v_q(t)}{\sqrt{\sum_{t \in T} v_p(t)^2} \cdot \sqrt{\sum_{t \in T} v_q(t)^2}}$$

$$s(x, y) = \frac{2 \cdot 3}{\sqrt{2^2 + 3^2}\sqrt{3^2 + 3^2}}$$

Data Integration: Data Matching

Table X

	Name	Phone	City	State
x_1	Dave Smith	(608) 395 9462	Madison	WI
x_2	Joe Wilson	(408) 123 4265	San Jose	CA
x_3	Dan Smith	(608) 256 1212	Middleton	WI

(a)

Table Y

	Name	Phone	City	State
y_1	David D. Smith	395 9426	Madison	WI
y_2	Daniel W. Smith	256 1212	Madison	WI

(b)

Matches
(x_1, y_1)
(x_3, y_2)

(c)

Figures from chapter 7 of "Principles of Data Integration"

- Data Matching is challenging due to variations in
  - formatting conventions
  - use of abbreviations, shortening
  - different naming conventions,
  - omissions
  - errors

    ⋮

Data Integration: Data Matching

Table X

	Name	Phone	City	State
x_1	Dave Smith	(608) 395 9462	Madison	WI
x_2	Joe Wilson	(408) 123 4265	San Jose	CA
x_3	Dan Smith	(608) 256 1212	Middleton	WI

(a)

Table Y

	Name	Phone	City	State
y_1	David D. Smith	395 9426	Madison	WI
y_2	Daniel W. Smith	256 1212	Madison	WI

(b)

Matches
(x_1, y_1)
(x_3, y_2)

(c)

Figure [illegible] from chapter 7 of “Principles of Data Integration”

- Methods
  - Rules-based method
  - Learning-based methods
    - Supervised learning
    - Clustering
    - probabilistic approach

Data Matching: Rule-Based

Table X

	Name	Phone	City	State
x_1	Dave Smith	(608) 395 9462	Madison	WI
x_2	Joe Wilson	(408) 123 4265	San Jose	CA
x_3	Dan Smith	(608) 256 1212	Middleton	WI

(a)

Table Y

	Name	Phone	City	State
y_1	David D. Smith	395 9426	Madison	WI
y_2	Daniel W. Smith	256 1212	Madison	WI

(b)

Matches

(x_1, y_1)
(x_3, y_2)

(c)

Figure [illegible] from chapter 7 of "Principles of Data Integration"

- a linearly weighted combination of the individual similarity scores between x and y:

$$sim(x, y) = \sum_{i=1}^{n} \alpha_i sim_i(x, y)$$

- A rule for the example in the figure

$$sim(x, y) = 0.3 s_{name}(x, y), +0.3 s_{phone}(x, y), +0.1 s_{city}(x, y), +0.3 s_{state}(x, y)$$

Data Matching: Rule-Based

Table X

	Name	Phone	City	State
x_1	Dave Smith	(608) 395 9462	Madison	WI
x_2	Joe Wilson	(408) 123 4265	San Jose	CA
x_3	Dan Smith	(608) 256 1212	Middleton	WI

(a)

Table Y

	Name	Phone	City	State
y_1	David D. Smith	395 9426	Madison	WI
y_2	Daniel W. Smith	256 1212	Madison	WI

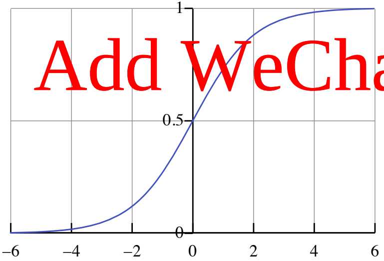
(b)

Matches

(x_1, y_1)
(x_3, y_2)

(c)

Figure is from chapter 7 of "Principles of Data Integration"



$$sim(x, y) = \frac{1}{1 + e^{-z}}$$

where

$$z = -\sum_{i}^{n} \alpha_i sim_i(x, y)$$

Figure is from chapter 7 of "Principles of Data Integration"

Data Matching: Learning-Based

- Supervised learning: learn a matching model with training data

$$T = \{(x_1, y_1, l_1), (x_2, y_2, l_2), \ldots, (x_n, y_n, l_n)\}$$

where (x_i, y_i) indicates a tuple pair, and l_i indicates the boolean label.

  - Define a set of features $f_1, f_2, \ldots, f_m$
  - Convert each training sample (x_i, y_i, l_i) into a feature vector

$$(< f_1(x_i, y_i), f_2(x_i, y_i), \ldots, f_m(x_i, y_i) >, c_i)$$

  - Apply supervised learning algorithms

Data Matching: Learning-Based

- Supervised learning: learn a matching model with training data

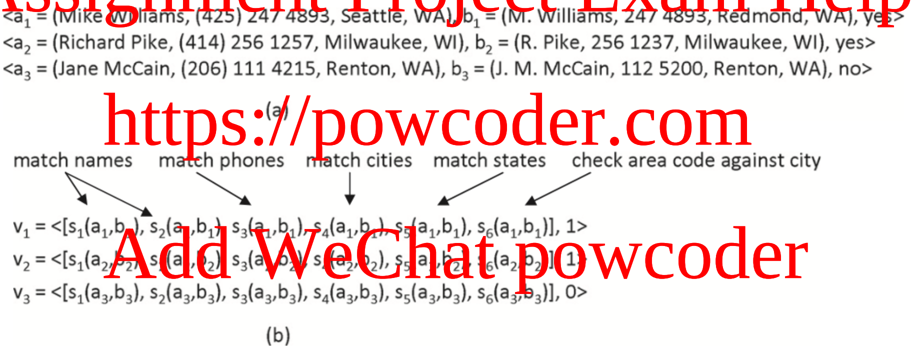


Figure is from chapter 7 of "Principles of Data Integration"

Data Matching: Learning-Based

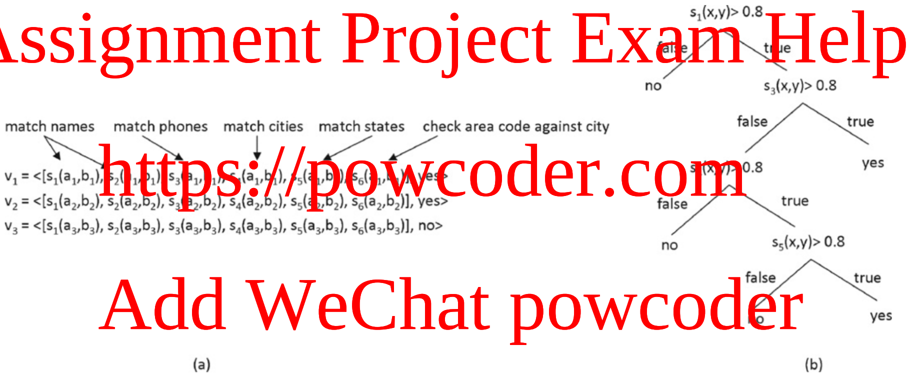


Figure is from chapter 7 of “Principles of Data Integration”

Data Matching: Learning-Based

MONASH University

- Clustering approach: tuples in the same cluster match
  - the problem of constructing entities(that is, clusters): only tuples within a cluster match.
  - An iterative process: leverage what we have known so far (in the previous iterations) to build "better" entities.
  - Generating a canonical tuple: "merge" all matching tuples within each cluster to construct an "entity profile".

Data Matching: Learning-Based

- Clustering approach: tuples in the same cluster match

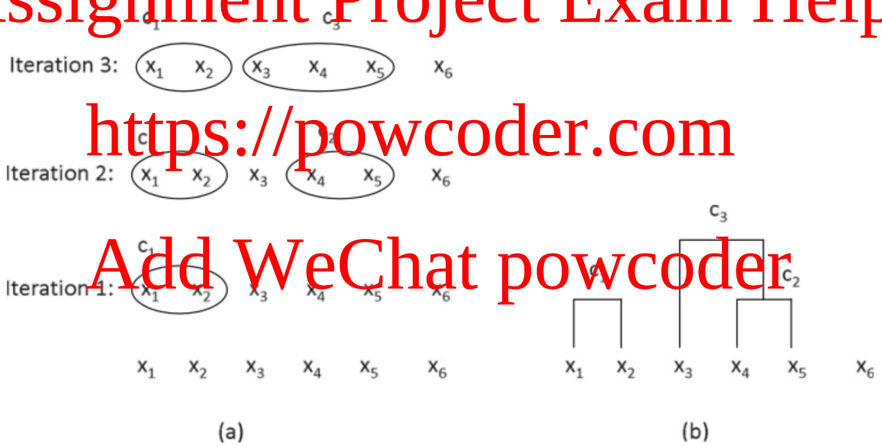


Figure is from chapter 7 of "Principles of Data Integration"

Summary

- Recap of schema integration
- Data integration: instance level
  - Attribute level integration
  - Tuple level integration
- Readings
  - Chapters 4 and 7, "Principles of Data Integration"
  - Chapter 5, "Data Matching-Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection"