

## Identifying Value Mappings for Data Integration: An Unsupervised Approach

Jaewoo Kang  
North Carolina State Univ.  
kang@csc.ncsu.edu

Dongwon Lee, Prasenjit Mitra  
The Pennsylvania State Univ.  
{dlee,pmitra}@ist.psu.edu

## Integrating Web Sources

- The Web has become the de facto standard for information dissemination and exchange.
- Useful information often scattered over multiple sites.
- Information integration across diverse sources is essential.

## Integrating Web Sources (II)

- Involves two subtasks:
  - Schema matching – resolving structural heterogeneity.
  - Object mapping – finding duplicate objects across data sources. (a.k.a., record linkage, deduplication)

## Previous Solutions

- Virtually all previous solutions assume data values in corresponding columns are drawn from the same domain, or at least share some textual similarity.
- This assumption is often violated in practice. E.g., “Two-door front wheel drive”, “2DR-FWD”, or “Car Type 3”.

## Value Mapping Problem

- Mapping data values across sources that represent the same real world concept.
- In particular, we focus on a difficult case where data values are opaque, or difficult to understand.
- Key Idea: we exploit not the tokens representing the data values but the *dependency structures* existing among data values that may characterize their semantic relations.

## Running example

Name	Gender	Title	Degree	Marital Status
J. Smith	M	Professor	Ph.D.	Married
R. Smith	F	Teaching Assistant	B.S.	Single
B. Jones	F	Teaching Assistant	M.S.	Married
T. Hanks	M	Professor	Ph.D.	Married

### University A

Name	Gender	Title	Degree	Marital Status
S. Smith	F	Emp10	Ph.D.	SGL
T. Davis	M	Emp3	M.S.	SGL
R. King	M	Emp10	Ph.D.	MRD
A. Jobs	F	Emp3	B.S.	MRD

## Running example

Name	Gender	Title	Degree	Marital Status
J. Smith	M	Professor	Ph.D.	Married
R. Smith	F	Teaching Assistant	B.S.	Single
B. Jones	F	Teaching Assistant	M.S.	Married
T. Hanks	M	Professor	Ph.D.	Married

### University A

Name	Gender	Title	Degree	Marital Status
S. Smith	F	Emp10	D7	SGL
T. Davis	M	Emp3	D3	SGL
R. King	M	Emp10	D7	MRD
A. Jobs	F	Emp3	D2	MRD

## The Algorithm

- $G_1 = \text{Table2CooccurrenceModel}(S_1);$   
 $G_2 = \text{Table2CooccurrenceModel}(S_2);$
- $\{(G_1(a), G_2(b))\} = \text{ModelMatch}(G_1, G_2);$

where  $S_i$  = an input table,  
 $G_i$  = a co-occurrence graph,  
 $(G_1(a), G_2(b))$  = a matching node pair.

## Step 1: Modeling Co-occurrence Relation

- Co-occurrence Matrix Model (CMM)
- Latent Semantic Model (LSM)

## Co-occurrence Matrix Model

	row <sub>1</sub>	row <sub>2</sub>	row <sub>3</sub>	row <sub>4</sub>		v1	v2	v3	v4	v5	v6	v7	v8	v9
v1:M	1	0	0	1	v1	2	0	2	0	2	0	0	2	0
v2:F	0	1	1	0	v2	0	2	0	2	0	1	1	1	1
v3:Professor	1	0	0	1	v3	2	0	2	0	2	0	0	2	0
v4:TA	0	1	1	0	v4	0	2	0	2	0	1	1	1	1
v5:Ph.D	1	0	0	1	v5	2	0	2	0	2	0	0	2	0
v6:M.S	0	0	1	0	v6	0	1	0	1	0	1	0	1	0
v7:B.S	0	1	0	0	v7	0	1	0	1	0	0	1	0	1
v8:Married	1	0	1	1	v8	2	1	2	1	2	1	0	3	0
v9:Single	0	1	0	0	v9	0	1	0	1	0	0	1	0	1

(a) Value-Row Matrix  $T_1$

(b) Co-occurrence Matrix  $C_1$

$$C_1 = T_1 * T_1^T$$

## Weighting Terms by Information Content

- Rare terms carry more weights when they co-occur than terms that occur frequently.
- We used standard Inverse Document Frequency (IDF) weighting as
  - $T_{ij} = 1 - k/N$  if term  $i$  occurs in row  $j$ , otherwise 0.
  - $k$  is the #of rows where term  $i$  occur and  $N$  is the total #of rows.

## Step 2: Find the mapping that

- minimizes the distance between the two Value Co-occurrence Matrices

## Capturing Indirect Relation

- Problem of the previous approach: strict pairwise co-occurrence!
- "A co-occurs frequently with B" and "B co-occurs frequently with C" → A and C may be relevant
- But the model does not capture this indirect co-occurrence relation if A never co-occurs directly with C in the data.

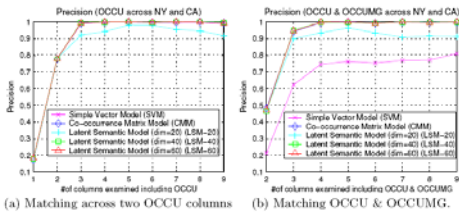
## Latent Semantic Model

- Like in Latent Semantic Indexing, using Singular Value Decomposition,
 
$$T1 = U * S * V^T$$

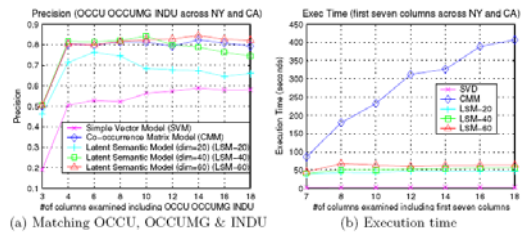
$$M1 = T1^k * (T1^k)^T$$
 Where  $T1^k$  is a best rank-k approximation of  $T1$  (obtained by keeping only top-k singular values and eigen vectors)
- Unlike  $C1$ ,  $M1$  captures the latent semantic relations.

## Experimental Validation

- Datasets: NY and CA census tables from U.S. Census Bureau.



## Experimental Validation (II)



## Conclusion

- Formally introduce value mapping problem.
- Proposed a novel algorithm that works for difficult cases where data values are opaque.
- The technique is invariant to the actual tokens representing the data, and so the technique is applicable to many different domains without training or tuning.
- Useful addition to the existing collection of data integration algorithms.