

Establishing Value Mappings Using Statistical Models and User Feedback

Jaewoo Kang, Tae Sik Han
North Carolina State Univ.
{kang,tshan}@csc.ncsu.edu

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

Value Mapping Problem

- Finding correspondence between data values across tables
- Limited study, only in the context of object mapping, record linkage, or approximate join.

2

Motivation

- Virtually all existing techniques assume values from corresponding columns are drawn from the same domain, or at least bear some textual similarity.
- This assumption often challenged in practice when heterogeneous tables being integrated.

3

Motivation (II)

- For example, "two-door front wheel drive" can be represented as "2DR-FWD" or "R2FD", or even as "CAR TYPE 3" in different data sources.
- Some smart string distance algorithms may be able to find correspondences among the first three representations, but they will fail to establish any mapping for "CAR TYPE 3".

4

Overview of Algorithm

- We propose iterative and interactive algorithm that does not rely on the tokens representing data.
- Instead, we exploit statistical dependencies among data values to find matches.
- The matches are further refined in each iteration by incorporating user feedback.

5

Overview of Algorithm (II)

- The algorithm works in three phases:
 - **Building** models independently for each table that characterize co-occurrence relations among data values.
 - **Matching** models across tables to find correspondence among data values.
 - **Refining** the models after each iteration incorporating user feedback.

6

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

University B

Running Example (II)

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

University B

Modeling Statistical Dependencies

- Any appropriate statistical models can be used for **Build** phase.
- We tested two alternative models: 1) co-occurrence frequency vector model, and 2) entropy vector model.

9

Co-occurrence Frequency Vector Model

- For each unknown value, v , build a signature vector
 - $\text{sig}(v) = [f(v), f(v, v_1), f(v, v_2), \dots, f(v, v_m)]$
 - where $f(\cdot)$ represents co-occurrence frequency of the two values, and $v_1 \dots v_m$ represent the values with known mapping. The first value $f(v)$ represents the frequency of v itself.

10

Co-occurrence Frequency Vector Model (II)

- Suppose we want to find mapping for "Professor".
- The signature vectors for "Professor" and the two values in the corresponding column, "Emp10" and "Emp3", are

$$\text{sig}(\text{Professor}) = [0.5], \text{sig}(\text{Emp10}) = [0.5], \text{and } \text{sig}(\text{Emp3}) = [0.5]$$
 initially when no other mappings are known.
- Suppose in the next iteration, "ph.D. -> D7" has been confirmed by user. Then, the vectors will be augmented with new frequency values as follows.

$$\begin{aligned} \text{sig}(\text{Professor}) &= [0.5, 0.5] \\ \text{sig}(\text{Emp10}) &= [0.5, 0.5] \\ \text{sig}(\text{Emp3}) &= [0.5, 0] \end{aligned}$$

11

Co-occurrence Frequency Vector Model (III)

- In practice, each frequency value, $f(a,b)$, can be weighted using inverse term frequency weight such as $(1-k/n) \cdot (1-l/n)$ where k and l are the numbers of times terms a and b occur in the table, and n is the total number of rows.
- Intuition is that we give more weights to the co-occurrence of rare values than the co-occurrence of common values.

12

Entropy Vector Model

- Uses entropy values conditioned with known mappings. E.g., when value in column X is fixed to x, entropy value for column Y is

$$H(Y | X = x) = - \sum_{y \in Y} p(y | x) \log p(y | x)$$

- Similarly, if values for two columns, X and Y, are fixed to x and y, entropy value for column Z is

$$H(Z | X = x, Y = y) = - \sum_{z \in Z} p(z | x, y) \log p(z | x, y)$$

13

Entropy Vector Model (II)

- E.g., signature vector for "Professor" in entropy model is $\text{sig}(\text{Professor}) = [H(\text{Gender} | \text{Title} = \text{"Professor"}), H(\text{Degree} | \text{Title} = \text{"Professor"}), H(\text{Marital} | \text{Title} = \text{"Professor"})]$ initially.
- Then, after "Ph.D." -> "D7" confirmed, $\text{sig}(\text{Professor}) = [H(\text{Gender} | \text{Title} = \text{"Professor"}), H(\text{Degree} | \text{Title} = \text{"Professor"}), H(\text{Marital} | \text{Title} = \text{"Professor"}), H(\text{Gender} | \text{Title} = \text{"Professor"}, \text{Degree} = \text{"Ph.D."}), H(\text{Marital} | \text{Title} = \text{"Professor"}, \text{Degree} = \text{"Ph.D."})]$.

14

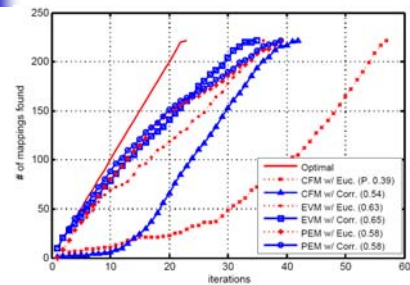
Experiments: Data Sets

No.	Col. name	# of unique values			Entropies		Description
		CA	NY	Common	CA	NY	
1	HHOUSUT	5	6	3	0.1522	0.1802	household type
2	HETENURE	3	3	3	1.0488	1.0558	own / rent / null
3	BRNUMHOU	13	13	12	2.8451	2.697	# of household members
4	HUFAMINC	17	16	16	3.6598	3.5568	total family income
5	PEEDUCA	17	17	17	3.2893	3.2759	highest degree earned
6	PEMARRTL	7	7	7	2.1347	2.227	marital status
7	PERACE	4	4	4	0.9873	1.0149	sex
8	PESEX	2	2	2	1	0.9971	sex
9	PVAFEVER	5	5	5	1.2143	1.1532	army veteran
10	PEMLR	8	8	8	2.1884	2.25	employment status
11	PRFTLF	4	4	4	1.8322	1.8075	full time, part time, etc
12	PERHCOW	9	8	8	1.7832	1.8056	class of worker (fed., priv., etc.)
13	PRDTIND1	49	49	47	3.4289	3.3758	industry code
14	PRDTOCC1	45	45	45	3.4792	3.4521	occupation code
15	PRMJIND1	23	22	22	3.0251	3.0193	industry - major group
16	PRMDOCC1	14	14	14	2.6512	2.664	occupation - major group
17	PEERNLAB	3	3	3	0.5211	0.5754	union member (y/n/null)

Table 3: Census table summary.

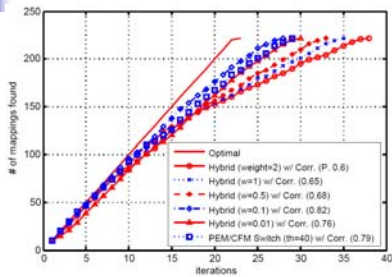
15

Co-occurrence Frequency Model vs. Entropy Vector Model



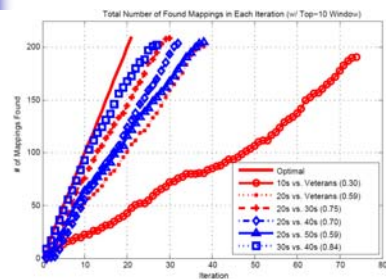
16

Hybrid Models



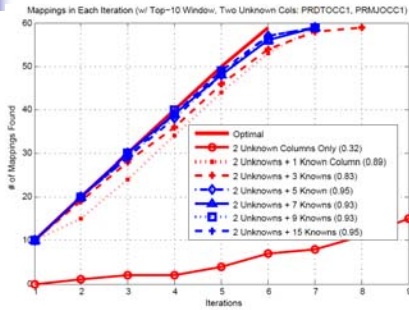
17

Sensitivity of Algorithm against Different Data Distributions



18

Effects of Pre-established Mappings



Conclusion

- Identified new class of value mapping problems that have not been addressed by existing solutions.
- Proposed a pragmatic solution that improves the mapping through iteration while incorporating user feedback.
- Evaluation suggests that it may be a useful addition to existing mapping techniques.