Mining the Heterogeneous Transformations for Record Linkage

Matthew Michelson & Craig A. Knoblock

Fetch Technologies   USC Information Sciences Institute

ICAI 2009

Record Linkage

Source 1

Manager	Restaurant
Bobby Jones	California Pizza Kitchen
William Smith	Arroyo Chop House
Bobby Smith	Panini Cafe

match

match

Source 2

Manager	Restaurant
Robert Jones	CPK
Bill Smith	Arroyo Steak Place
Bob Smith	The Pancake Palace

Heterogeneous Transformations

- Not characterized by a single function
  (vs. edit distances …)
  - Synonyms/Nicknames
    - Robert → Bobby
  - Acronyms
    - California Pizza Kitchen → CPK
  - Representations
    - 4^{th} → Fourth
  - Specificity
    - Los Angeles → Pasadena
  - Combinations
    - Sport Utility 4D → 4 Dr SUV

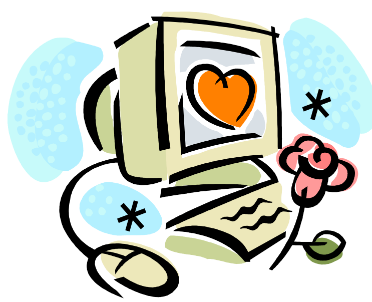
Heterogeneous Transformations

- Applications
  - Record linkage
    - Disambiguating records: Robert = Bobby
- Information retrieval
  - Search: “4dr SUV” Return: “4 door Sport Util…”
- Text understanding
  - Acronyms, Synonyms, Specificities
- Information extraction
  - Expand extraction types

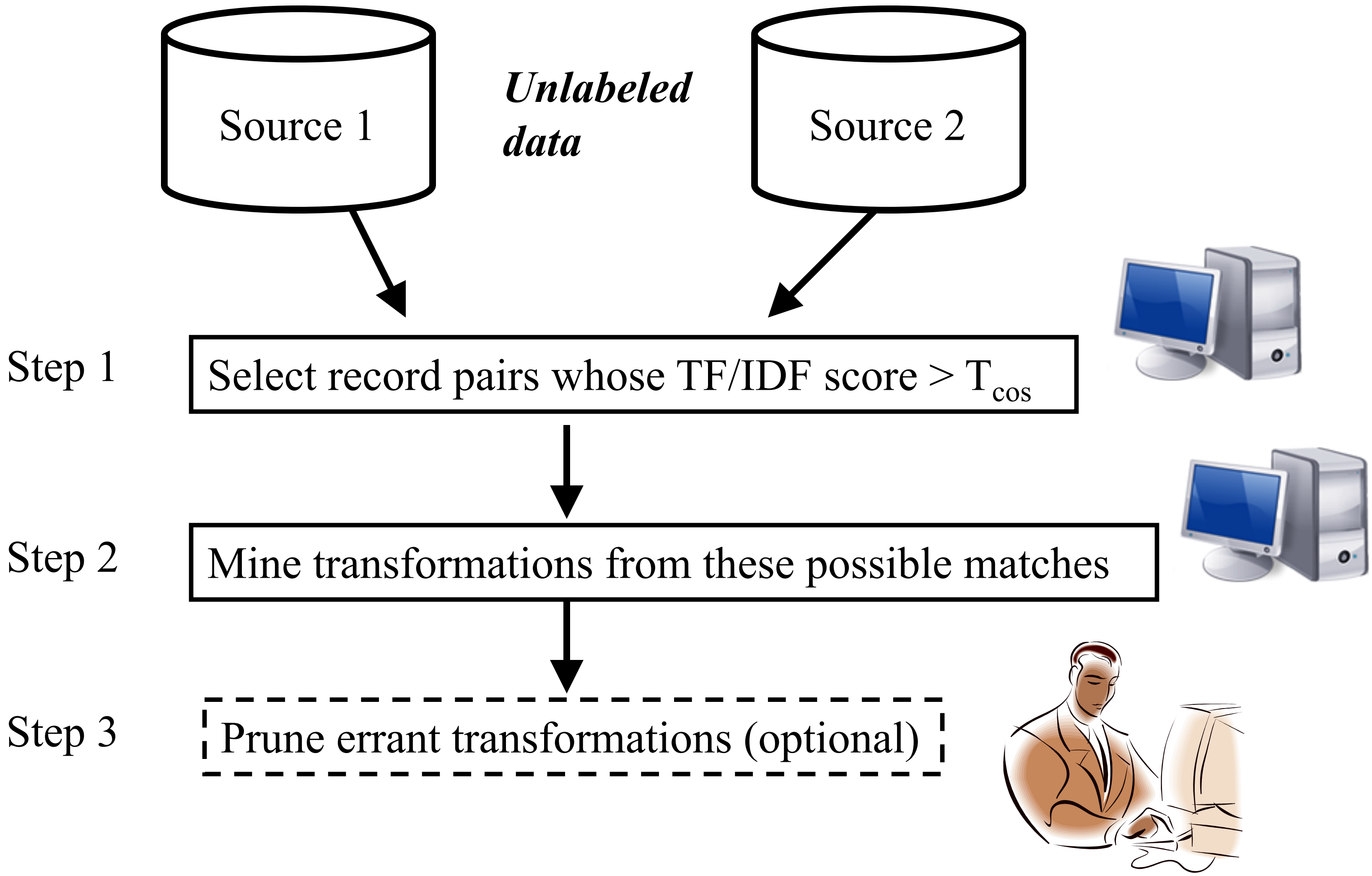
Heterogeneous Transformations

- *Before:* Manually created a priori

- *Now:* Mined from datasets,
  - minimal human effort



Algorithm overview (3 steps)



Step 1: Selecting record pairs

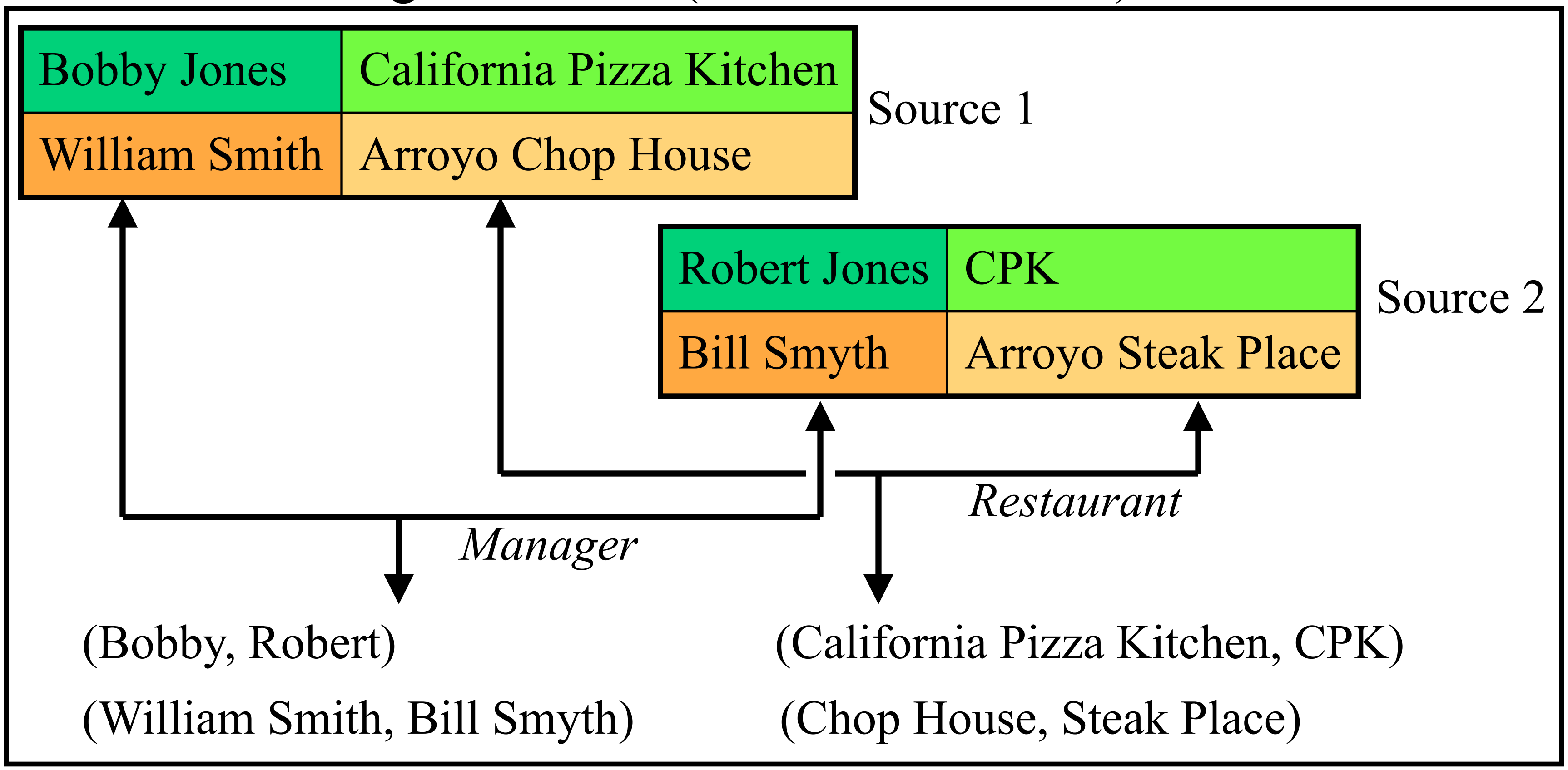
- Select record pairs that are "close"
  - High token-level simiarity
  - Loosens requirement on training data
  - "Close" is not exact
    - Share some similarity
    - Mine transformations from differences

Bobby Jones	California Pizza Kitchen
William Smith	Arroyo Chop House

Robert Jones	CPK
Bill Smyth	Arroyo Steak Place

Step 2: Mining Transformations

1. Get co-occurring token sets (not exact matches)



2. Select token sets with mutual information $> T_{MI}$

Mutual Information

- $$MI(s,t) = p(s,t) * \log\left(\frac{p(s,t)}{p(s)p(t)}\right)$$
- high mutual information
  - occur together with a high likelihood
  - carry information about the transformation occurring in that field for possible matches

Results: Example Mined Transformations

Cars Domain		
Field	*Kelly Blue Book Value*	*Edmunds Trans.*
Trim	Coupe 2D	2 Dr Hatchback
Trim	Sport Utility 4D	4 Dr 4WD SUV *or* 4 Dr STD 4WD SUV *or* 4 Dr SUV
BiddingForTravel domain		
Field	*Text Value*	*Hotel Trans.*
Local area	DT	Downtown
Hotel name	Hol	Holiday
Local area	Pittsburgh	PIT (airport code!)
Restaurants domain		
Field	*Fodors Value*	*Zagats Trans.*
City	Los Angeles	Pasadena *or* Studio City *or* W. Hollywood
Cuisine	Asian	Chinese *or* Japanese *or* Thai *or* Indian *or* Seafood
Address	4th	Fourth
Name	and	&
Name	delicatessen	delis *or* deli

Results: Threshold Behavior

- More sensitive to T_{MI} than T_{cos}
  - T_{MI} picks transformations, T_{cos} picks candidate matches
- Lower T_{MI} yields more transformations
  - Fewer transformations are common ones

  → bad discriminators for record linkage (e.g. 2dr = 2 Door)
- Setting T_{cos} too high limits what can be mined
- Strategy
  - Set Tcos low enough so it's not too restrictive
  - Set TMI low enough so that you mine a fair number of transformations
    - Yields noise, but does not affect record linkage

Results: Record Linkage Improvement

RL experiments use $T_{cos} = 0.65$ and $T_{MI} = 0.025$, for threshold sensitivity results, see paper

	Recall	Prec.
Cars domain		
No trans.	66.75	84.74
Full trans.	**75.12**	83.73
Pruned trans.	75.12	83.73
BFT domain		
No trans.	79.17	93.82
Full trans.	**82.89**	92.56
Pruned trans.	82.47	92.87
Restaurants domain		
No trans.	91.00	97.05
Full trans.	91.01	97.79
Pruned trans.	90.83	97.79

In **all** domains, not stat. sig. between pruned set & full set → *pruning optional*

Trans. mostly in "cuisine" but decision tree ignores this field

Conclusions and Future Work

- Conclusions:
  - Mine transformations without labeling data
  - Pruning errant transformations is optional
- Future Work
  - Some fields are ignored, so waste time mining
    - Predictable?
  - Better candidate generation
    - Different methods?
  - Explore technique with other applications

Related Work

- Similar to association rules (Agrawal, et. al. 1993)
  - Even mined using mutual information (Sy 2003)
  - Assoc. rules defined over set of transactions
    - "users who buy cereal also buy milk"
  - Our transformations defined between sources
- Phrase co-occurrence in NLP
  - IR results to find synonyms (Turney 2001)
  - Identify paraphrases & generate grammatical sentences (Pang, Knight & Marcu 2003)
  - We are not limited word based transformations: "4d" is "4 Dr"
    - No syntax is needed

Thank you!