Entity Resolution

Yibin Xiong

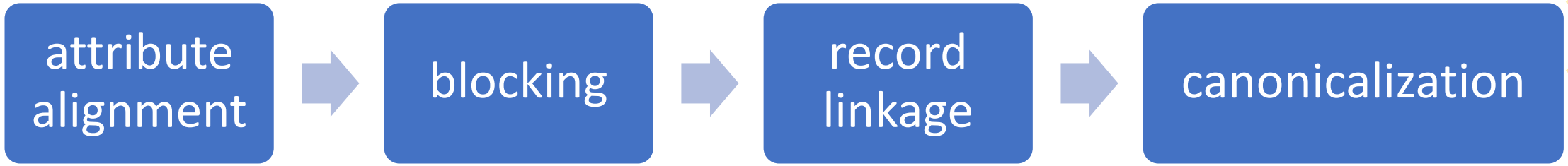
Problem to Solve

- Matching different **records** (rows) to the same entity
  - Within a database: **de-duplication**
  - Across multiple databases: **record linkage**
- Unique identifiers are unavailable
- Inaccurate and missing data

Record	Given name	Family name	Year	Month	Day	Municipality
1.	JOSE	FLORES	1981	1	29	A
2.	JOSE	FLORES	1981	2	NA	A
3.	JOSE	FLORES	1981	3	20	A
4.	JULIAN ANDRES	RAMOS ROJAS	1986	8	5	B
5.	JILIAM	RMAOS	1986	8	5	B

Table 2: Illustrative example of duplicated records in the UNTC data set reproduced from Table 1 of Sadinle (2014). Note that record 5 most likely has errors, where "RMAOS" should be "RAMOS," due to the processing of list photocopies using Optical Character Recognition. These errors were corrected in Sadinle (2014).

Pipeline



- Attribute alignment: find the common attributes across databases
- Blocking: propose possibly matching pairs (rough filtering)
- *Record linkage: classify if candidate pairs match or not (fine filtering)
- Canonicalization: merge matched records to a single representative record

Blocking: Simple, Deterministic Rules

- Determine blocking key from one or several attributes
  - E.g. name; gender AND date of birth
- (Optionally,) Map the attribute values to blocking key values (BKVs)
  - E.g. phonetic encoding functions map names with similar sound to the same code
- Group the records with the same BKV together

Table 4.1 Example name strings and their phonetic encodings. Variations of the same name are grouped together

String	Soundex	Phonex	Phonix	NYSIIS	Double Metaphone	Fuzzy Soundex
peter	p360	b360	p300	pata	ptr	p360
pete	p300	b300	p300	pat	pt	p300
pedro	p360	b360	p360	padr	ptr	p360
stephen	s315	s315	s375	staf	stfn	s315
steve	s310	s310	s370	staf	stf	s310
smith	s530	s530	s530	snat	sm0, xmt	s530
smythe	s530	s530	s530	snat	sm0, xmt	s530
gail	g400	g400	g400	gal	kl	g400
gayle	g400	g400	g400	gal	kl	g400
christine	c623	c623	k683	chra	krst	k693
christina	c623	c623	k683	chra	krst	k693
kristina	k623	c623	k683	cras	krst	k693

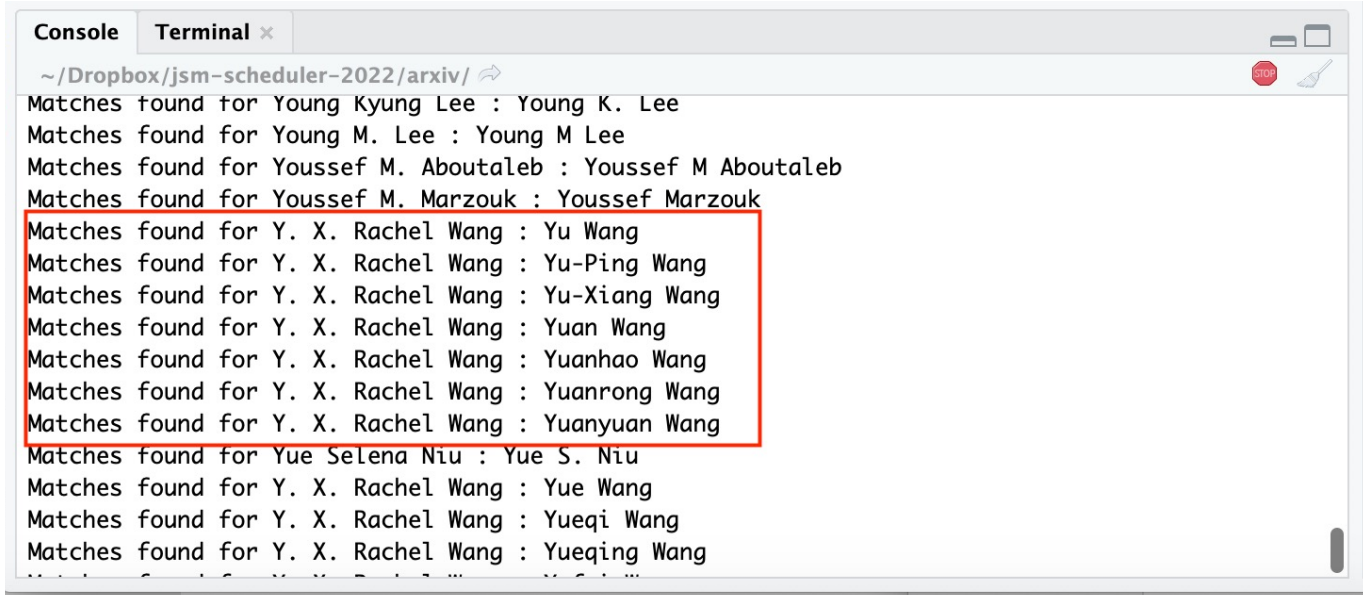
Id	Last Name	First Name	Postal code
1	Smith	Anna	1234 AB
2	Smith	George	1234 AB
3	Schwarz	Ben	6789 XY

Id	Last Name	First Name	Postal code
1	Smith	George	1234 AB
2	Johnson	Charles	1234 AB
3	Schwarz	Ben	6789 XY
4	Schwarz	Anna	6789 XY

A_Id	B_Id	Postal code
1	1	1234 AB
1	2	1234 AB
2	1	1234 AB
2	2	1234AB
…	…	…

Blocking

- For deduplicating arXiv, we use the rule of:
  - The two names must be *compatible* (i.e. last names exactly match, initials of first names match)
  - The two records share at least 1 co-author
- Problem: Too many false discoveries if no further filtering



Record Linkage: Probabilistic Rules [Fellegi & Sunter 1969]

- Let $\gamma = (\gamma_1, \ldots, \gamma_k)$ be the comparison vector between two records
- Estimate $\mathrm{m}(\gamma) \coloneqq \mathbb{P}(\gamma \mid M)$ and $\mathrm{u}(\gamma) \coloneqq \mathbb{P}(\gamma \mid U)$
  - Supervised learning: estimate m,u probabilities from training data
  - *Unsupervised learning: EM algorithm
- Consider the log ratio $W(\gamma) = \log m(\gamma) - \log u(\gamma)$
- Define constants T_μ, T_λ for controlling the Type I errors μ, λ
  - Match if $W(\gamma) > T_\mu$
  - Undetermined if $T_\lambda < W(\gamma) \leq T_\mu$
  - Non-match if $W(\gamma) \leq T_\lambda$

Theoretical Properties of F&S

- Essentially, Fellegi & Sunter method is a likelihood ratio test
- The rule is the *optimal* one in the sense of minimizing the probability of a comparison vector being undetermined
- Bound the ratio $\frac{m(\gamma)}{u(\gamma)}$ is equivalent to bound the posterior probability

$$\mathbb{P}(M \mid \gamma) = \frac{\mathbb{P}(M)m(\gamma)}{\mathbb{P}(M)m(\gamma) + \big(1 - \mathbb{P}(M)\big)u(\gamma)} = 1 - \left(1 + \frac{\mathbb{P}(M)}{1 - \mathbb{P}(M)} \cdot \frac{m(\gamma)}{u(\gamma)}\right)^{-1}$$

i.e. for very large/small likelihood ratio, the posterior probability is also large/small, in which case we reject the null hypothesis and classify the pair as match/non-match

Typical Model Settings

- Binary comparison vector γ [Enamorado 2018]
  - If an attribute is *string*-valued, use *edit distances* such as Levenshtein, Jaro, and Jaro-Winkler distance and convert it to similarity.
  - If an attribute is numerical, use L1 or L2 distance
  - $\gamma_k(i,j) = \begin{cases} 1, & dist_k(i,j) < \tau_k \\ 0, & dist_k(i,j) \geq \tau_k \end{cases}$
- Categorical comparison vector γ [Enamorado et al. 2019]
  - Measure distance/similarity and ***discretize*** it into L_k bins
- Models for probability distributions [Enamorado 2018]

$$\gamma_k(i,j) \mid M(i,j) = m \overset{\text{indep.}}{\sim} \text{Discrete}(\boldsymbol{\pi}_{km})$$

$$M(i,j) \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(\lambda)$$

- M(i,j) is a latent variable. We can use EM algorithm.

Implementation

- Tasks
  - De-duplicate ArXiv
  - De-duplicate Semantic Scholar(S2)
  - Match S2 to JSM
  - Match ArXiv to JSM
- Data
  - JSM: author-coauthor (5274*5274), author-word (5274*3966)
    - JSM authors are unique; Data are used for matching
  - ArXiv: paper-author (90434*113507), author-coauthor (113507*113507), *author-word
  - S2: author-paper (326895*280158), author-coauthor, author-word
- Attributes for record linkage (They are integer-valued, so already *discretized*)
  - De-duplication: name similarity (or frequency measure), # of common co-authors
  - Matching: name similarity (for blocking), word usage similarity (We can't use # of common co-authors for matching because the scopes of different author-coauthor matrices are different.)
  - I also want to use affiliation, but the available affiliation in S2 is too scarce

*Need to additionally download titles and/or abstracts

Experiments

- Vanilla (what we did in the summer)
- RL (record linkage) + TF-IDF
- RL + word embedding downloaded from S2 (If time permits)

Software

- fastLink: https://github.com/kosukeimai/fastLink
- RecordLinkage: http://uribo.github.io/rpkg_showcase/modeling/RecordLinkage.html
- *reclin (good demo, flexible to use): https://github.com/djvanderlaan/reclin

References

- Binette, Olivier, and Rebecca C. Steorts. "(Almost) All of Entity Resolution." arXiv e-prints (2020): arXiv-2008.
- Enamorado, Ted, Benjamin Fifield, and Kosuke Imai. "Using a probabilistic model to assist merging of large-scale administrative records." American Political Science Review 113.2 (2019): 353-371.
- Enamorado, Ted. "Active learning for probabilistic record linkage." Available at SSRN 3257638 (2018).