Eric & Wendy Schmidt

Data Science For Social Good

Summer Fellowship

Record Linkage / Matching

THE UNIVERSITY OF CHICAGO

Goals

- Determine if pairs of *records* describe the same entity

- Main applications:
  - *Joining* two different data sources
  - *Removing duplicates* from a single data source

Synonyms (pun intended)

- data matching
- merge/purge
- duplicate detection
- de-duping
- reference matching
- co-reference/anaphora resolution

Factors to consider

- Deduping or Linkage
  - 1-1 or 1-many or many-1
- Rule-based or Probabilistic
  - Do you have labeled training data?
- Domain specific or generic similarity functions?
- Evaluation metric
  - Precision or recall
  - Implications on future analysis

Approaches

- Exact matching
- Rule-based
- Probabilistic linkage

Common mismatches

- Case
- Nicknames
- Prefixes
- Suffixes
- Initials
- Punctuation
- Spaces
- Digits
- Transpositions
- Abbreviations

Common distance metrics

- Edit distance
- Soundex

“Fuzzy” Matching System

- Apply set of cascading rules
- Assign confidence score based on which rules fire

Efficiency

- How do we avoid looking at |A| * |B| pairs?
- *Blocking:* choose a smaller set of pairs that will contain all or most matches.
  - Simple blocking: compare all pairs that "hash" to the same value (e.g., same Soundex code for last name, same birth year)
  - Extensions (to increase *recall* of set of pairs):
    - Block on *multiple* attributes (soundex, zip code) and take union of all pairs found.
    - *Windowing:* Pick (numerically or lexically) *ordered* attributes and sort (e.g., sort on last name). The pick all pairs that appear "near" each other in the sorted order.

Machine Learning based Record Linkage

- Generate training data
  - Label pairs as match/no match
- Generate features over each pair
  - Distance metrics over different attributes (fname, lname, dob, etc.)
  - Tfidf scores
- Build classifiers

Tools

- Lots of commercial tools
  - IBM (good – used to be initiate systems)
  - Dataladder
  - …
- Open source
  - DeDupe
  - FRIL: http://fril.sourceforge.net/
  - …