Adaptive Sorted Neighborhood Methods for Efficient Record Linkage

The Pennsylvania State University #
National University of Singapore *

# Su Yan

Record Linkage (RL)

- Finding entries that refer to the same entity
- A well-studied area
  - Citation matching – digital library
  - Merge-purge – database
  - Duplicate detection, entity resolution …
- Datasets (databases) are usually “dirty”
  - Data from multiple sources
  - Spelling mistakes…
- RL as a pre-requisite step for data mining
  - Garbage-in, Garbage-out

RL Steps

- Why blocking?
  - Prevent pair-wise comparisons
  - Fast, approximate method to generate candidate matches
  - E.g.: First Initial + Last name

Limitations of RL

- Not very flexible
  - Employ parameters whose value are set by human experts
    - E.g.
      - Attributes to use for blocking
      - Choice of blocking algorithms
      - Choice of similarity functions
    - Finding the ideal values for such parameters is not straightforward

"Adaptivity" to the rescue

- Research Question
  - Whether a record linkage method can adaptively choose different parameters for different data sets and configurations with little human intervention?
- Existing work toward adaptive record linkage
  - Learn an optimal blocking strategy [Bilenko et al 2006] [Michelson et al 2006]
  - Learn string edit distance for different data sets [Bilenko et al 2003]
- Our goal
  - Find blocks whose sizes are adaptive to the duplicate distribution of the data set

Why adaptive blocking is important?

Ideal block sizes in the Cora data set
(1295 citations, 116 blocks, alphabetically ordered)
Sorted Neighborhood Method (SNM)
- classical method
- simple to implement
- widely used
- running time $O(wn)$

Adaptive SNMs
- Goal: Instead of the predetermined window size, dynamically adjust window sizes to suit the duplicate distribution of data set.
- Two adaptive versions of SNM
  - Incrementally-adaptive SNM (IA-SNM)
  - Accumulatively-adaptive SNM (AA-SNM)

Incrementally-adaptive SNM (IA-SNM)
- Basic idea: To find out if records in a window are close/sparse and if there are rooms to grow/shrink the window?

Accumulatively-adaptive SNM (AA-SNM)
- Basic idea: similar to how people watch videos – i.e., if subsequent scenes are similar, press fast-forward to skip frames to arrive at new scenes quickly, and press fast-backward to go back if too many frames are skipped.

Experiment set up
- Evaluation metrics
  - Pairs Completeness
  - Reduction Ratio
  - F-score
  - Baseline method
  - exact blocking
- Data sets

Experiments with real data
- Cora -- citation
- Restaurant -- address
Varying the error rate of duplicates

Influence of error rate to blocking schemes
(data set: mailing list, size: 10,000)

Varying the size of blocks

Conclusion and Future work

- Conclusion
  - Advocated the importance of adaptivity in record linkage problems;
  - Studied the adaptivity problem in detail using the classical SNM record linkage algorithm;
  - Presented two adaptive versions of the SNM algorithm and showed their efficacy in several dimensions;

- Future work
  - Apply the adaptivity idea to other existing blocking methods;
  - A comprehensive adaptive record linkage framework.

Thanks for your attendance!

Questions?