

Adaptive Sorted Neighborhood Methods for Efficient Record Linkage

The Pennsylvania State University #
National University of Singapore *

Su Yan
With # Dongwon Lee, * Min-Yen Kan, # C. Lee Giles

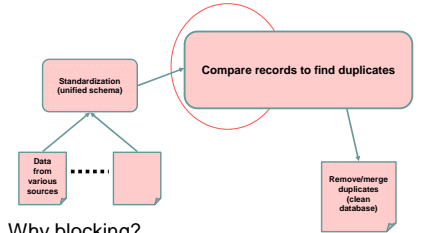


Record Linkage (RL)

- 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

JCDL 2007 2

RL Steps



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

JCDL 2007 3

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

JCDL 2007 4

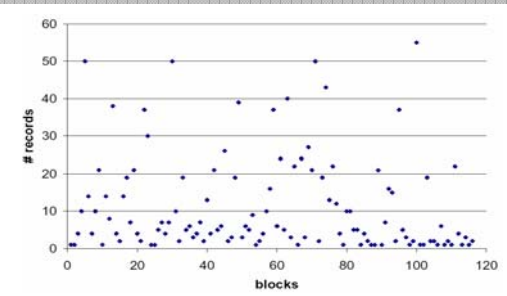
“Adaptivity” to the rescue

- 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

JCDL 2007 5

Why adaptive blocking is important?



Ideal block sizes in the Cora data set
(1295 citations, 116 blocks, alphabetically ordered)

JCDL 2007 6

Sorted Neighborhood Method (SNM)

- classical method
- simple to implement
- widely used
- running time $O(wn)$

Window size is fixed to 3

Report duplicate

JCDL 2007 7

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)

JCDL 2007 8

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?

JCDL 2007 9

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.

JCDL 2007 10

Experiment set up

- Evaluation metrics
 - Pairs Completeness $PC = \frac{\#Correctly\ Identified\ Duplicate\ Pairs}{\#True\ Duplicate\ Pairs}$
 - Reduction Ratio $RR = 1 - \frac{\#Identified\ Potential\ Duplicate\ Pairs}{\#All\ Pairs}$
 - F-score $F\text{-score} = \frac{2 * RR * PC}{RR + PC}$
- Baseline method
 - exact blocking
- Data sets

Summary of the data sets for experiments

Database name	size	property	#field	content	#blocks	maximum #records per block
Cora	1265	real	12	citation	116	55
Restaurant	864	real	4	restaurant addresses	112	2
DBCen	varies	synthetic	9	mailing list	varies	varies

JCDL 2007 11

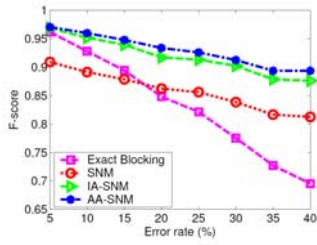
Experiments with real data

Cora -- citation

Restaurant -- address

JCDL 2007 12

Varying the error rate of duplicates

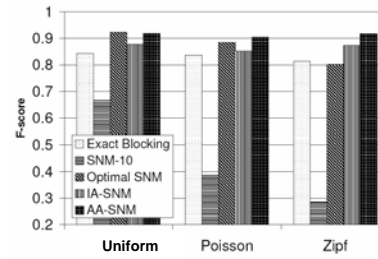


Influence of error rate to blocking schemes
(data set: mailing list, size: 10,000)

JCDL 2007

13

Varying the size of blocks



Mailing List dataset

JCDL 2007

14

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.

JCDL 2007

15

Thanks for your attendance!

Questions?

JCDL 2007

16