An Unsupervised Algorithm for Learning Blocking Schemes
Mayank Kejriwal1 and Daniel P. Miranker1
1 {kejriwalm, miranker}@cs.utexas.edu / The University of Texas at Austin

Unsupervised Deduplication

We enable it for the first time

Algorithm: Overview

Tabular dataset

Generate pseudo training set

Perform feature selection

DNF Blocking Scheme

Ongoing Work: Overview

- Build complete and unsupervised deduplication system (preliminary results available)
- Implement unsupervised deduplication system in MapReduce framework
- Experiment on Big Data (tens of gigabytes)
- More advanced feature selection (like Laplacian)
- Time-series and numerical data

Algorithm: Steps

- Treat each record as a bag of words and block on its tokens
- Slide a window of constant size $c$ over each block and calculate log TF-IDF scores of record pairs within each window
- If score is above a threshold $s$, record pair classified as duplicate
- If score is below a threshold $s$, record pair classified as non-duplicate
- Given noisy sets of duplicates $D$ and non-duplicates $N,D$, perform feature selection
- Optimal feature selection shown to be NP-hard in previous work
- Use greedy algorithm with Fisher criterion to form approximate solution
- Reverse map chosen features to a DNF blocking scheme

Experimental Results: Pseudo Training Set Generation

Three standard benchmarks used: Restaurant, Census and Cora

Precision of duplicates retrieved on Restaurant and Census

Precision of duplicates retrieved on Cora

Effects of varying upper threshold parameter $c$

Effects of varying lower threshold parameter $t$

Ongoing Work: Preliminary Results

We tested our completely unsupervised deduplication system on two synthetic datasets derived from real-world census data and one real-world benchmark used earlier in this paper

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Census (1000 records)</td>
<td>100.00</td>
<td>100.00</td>
</tr>
<tr>
<td>Census Noise (1000 records)</td>
<td>100.00</td>
<td>61.17</td>
</tr>
<tr>
<td>Restaurant (real world)</td>
<td>99.71</td>
<td>99.76</td>
</tr>
</tbody>
</table>

Learning a DNF Blocking Scheme

Learning Disjunctive Blocking Schemes

Learning DNF Blocking Schemes

References