Hash-based Similarity Detection

Related Work

- Locality-sensitive hashing (LSH)
  - Nilsimsa
  - SimHash
- Distance measure
  - Hamming distance

Method

- Column Representation
  - set-based
  - hash-based
- Column Comparison
  - set intersection
  - hash similarity (Hamming distance)
  - pairwise comparison: O(n^2) vs pre-ordering
- Table Comparison ?
  - pairwise column comparison: O(nm)
  - table representation: O(1)

Experiments

- Evaluate hash-based similarity **within** the column
  - hash table cells within the same column
  - sort by similarity
- Evaluate hash-based similarity **between** the columns
  - Table file: 'httpwww.wienticket.atfeedsvorverkauf.phpformatcsv'
  - same column ‘Ort’ split into two disjunct subsets
  - + 2 real sample columns ['Adresse', 'Vorverkaufstelle']
  - 1192 rows (596 - half)
  - SimHash: hashbits=8

Results (within the column)

Ernstbrunn

Hollabrunn

Poysdorf

Spillern

Mistelbach

Korneuburg

Wien

Wolkersdorf

Bisamberg

Stockerau

Wien-Flughafen

Results (between the columns)

Adresse Ort2

0.375

Adresse Vorverkaufstelle

0.5

Adresse Ort1

0.375

Ort2 Vorverkaufstelle

0.375

Ort2 Ort1

1.0

Vorverkaufstelle Ort1

0.375

Column1 Column2

Similarity [0:1]

Limitations

- Columns need to be of the same length to produce the same hashes
- Similarity threshold ?