ALIAS

An Active Learning led Interactive Deduplication System

Sunita Sarawagi  Anuradha Bhamidipaty
Alok Kirpal        Chandra Mouli
{sunita,anu,alok,mouli}@it.iitb.ac.in
The duplicate elimination problem

- Eliminate duplicates from lists of records
- Manual design of de-duplication function hard, as
  - data has errors and inconsistencies
  - duplicates may be spread far apart
- Solution: Learn function from examples
- Providing useful examples hard
- Solution: Discover using active learning
Architecture of ALIAS
Working of ALIAS

- Apply similarity functions on record pairs.
- Loop until user satisfaction
  - Train classifier.
  - Use active learning to select n instances
  - Collect user feedback.
  - Augment with pairs inferred using transitivity
  - Add to training set
- Output classifier
Active learning: selecting instances

- Train k classifiers $C_1, C_2, \ldots, C_k$ on training data.
- For each unlabeled instance $x$:
  - Find prediction $y_1, \ldots, y_k$ on the k classifiers.
  - Compute uncertainty $U(x)$ as entropy of $y$-s.
- With weight as $U(x)$, do weighted sampling to select an instance for labeling.
Benefits of active learning

Learning de-duplication function on Bibtex entries

With 100 pairs:
- Active learning: 97% (peak)
- Random: only 30%
Features of ALIAS

- Interactive discovery of deduplication function using active learning
- Manual effort reduced to
  - Providing simple similarity functions
  - Labeling selected pairs
- Efficient indexing mechanism
- Novel cluster-based evaluation engine
- Cost-based optimizer