Entity Resolution, Clustering Author References

Vlad Shchogolev
vs299@columbia.edu

May 1, 2007

Outline

1. Introduction
   - What is Entity Resolution?
   - Motivation
2. Background
   - Formal Definition
   - Efficieny Considerations
   - Measuring Text Similarity
   - Other approaches
3. Methodology
   - Clustering author references
   - Learning distance function
   - Clustering experiment

Set-Up

- Given: a table of records, each record has a number of "fields".
- Problem: decide which pairs of records refer to the same "entity".
- Variation: given two tables, pair up matching records

A new name for an old research area

- Record linkage
    - Originally studied by Dunn, 1946
    - Formalized by Fellegi and Sunter, 1969
- Merge/purge problem
- Data matching, object identity problem
- Coreference resolution, reference reconciliation, etc.

Why is this useful?

- Historical research
- Medical research
- Government record-keeping
- Tracking terrorists
- Wal-Mart
- Yahoo Local, Google Local
- Bibliographic citations

Formal Definition

- Two datasets A and B, and let $(a, b) \in A \times B$.
- Also define M as the set of matched pairs $(a = b)$ and U as the set of unmatched pairs.
- For a given pair, there is a list of comparison values for $(c_1, \ldots c_n)$ where c_k is the comparison value for the k^{th} field of the item.

Probabilistic Model, Fellegi and Sunter

- Define quantities $m_k = P\{c_k = 0 | (a, b) \in M\}$ and $u_k = P\{c_k = 0 | (a, b) \in U\}$
- Assume independence assumption (!)
- Weight assigned to each component is:

$$w_k = \begin{cases} \log(m_k/u_k) & \text{if } c_k = 0, \\ \log(1 - m_k)/(1 - u_k)) & \text{if } c_k = 1. \end{cases}$$

- Equivalent to Naive Bayes

Canopies

- Need methods to find candidate pairs
- McCallum, Nigam, Ungar introduced “canopies” approach: clustering is performed in two stages, starting with a rough stage that divides data into overlapping subsets
- If clustering measures distance to cluster using cluster centroid, and canopy is larger than true cluster, nothing is lost.

Clustering bibliographic references

- Goal was to compute the citation graph for research papers
- First pass: used a fast TF-IDF approach
- Second pass: used an expensive string edit distance computation, combined with a HMM for field extraction
- Results in equally good accuracy, but orders of magnitude faster

Measuring Text Similarity

Several methods exist to construct a similarity measure for text:

- edit distance (customizable costs)
- Jaro's algorithm (transpositions)
- character N-grams
- TF-IDF
- string kernels
- term-vector dot product
- soundex

Research typically finds that no single method is best

Refinements

Minton, Nanjo et al. introduced “transformation graphs” to handle higher level concepts:

- synonyms
- misspelling
- abbreviation
- acronym
- concatenation

Again, a Naive Bayes approach is used to learn weights for transformations.

Domain-independent approach

- Monge and Elkan suggest a “domain-independent” approach
- Each record is treat as a single long string
- Similarity is measured using edit distance

Clustering author references

- A related problem to bibliographic references
- Experiment run on a large biology research corpus
- Goal is to determine when two matching names (e.g. Smith J.) refer to the same person
- Extra structure: social network consisting of co-authorship edges

Learning distance function

- Similar to distance metric learning paper, but not in Euclidean space
- Not domain-independent
- Domain knowledge used to pick from one of 3 comparison functions for each field:
  - equality
  - set intersection
  - character N-gram similarity
- Most important feature: number of common co-authors

Clustering experiment

- Clustering name references can be useful for judging co-authorship importance
- Tried experiments with simple Greedy Agglomerative Clustering.
- Measured within-cluster dispersion at each clustering step:
  - Data is clustered into k clusters $C_1, \ldots C_k$
  - Sum of pairwise distances: $D_r = \sum_{i,j \in C_r} d_{i,j}$
  - Dispersion measure: $W_k = \sum_{r=1}^{k} \frac{1}{2n_r} D_r$

Gap Statistic

- Can we estimate the true number of clusters?
- Define Gap Statistic (Tibshirani et al. 2000)

$$Gap_n(k) = E_n^*(\log(W_k)) - \log(W_k)$$

- Expectation is over a sample from reference distribution
- The quantity $\log(W_k)$ can be thought of as log-likelihood

Reference distribution

- $Gap_n(k) = E_n^*(\log(W_k)) - \log(W_k)$
- For uniform distribution, expectation should decrease at the rate $(2/p)\log k$.
- Better: a uniform distribution over a box align with the principal components.
- Goal is to produce evidence against the null model (single cluster)

Estimation Procedure

- Sample B Monte Carlo reference datasets from reference distribution
- Cluster each sampled dataset
- Estimate the gap:

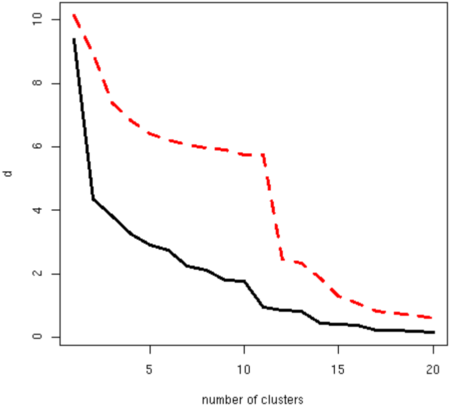
$$Gap(k) = (1/B)\sum_b \log(W^*_{kb}) - \log(W_k)$$

- Pick smallest k such that $Gap(k) > Gap(k+1) - s_{k+1}$

Estimation Procedure

- Estimation performs poorly; reference distribution not appropriate?
- Euclidean distance metric not appropriate for high dimensions
- However: have evidence that W_k graph has some signal

Two within-cluster dispersion graphs



Summary

- Rethink cluster estimation for this class of problems
- Distance metric learning works well, but...
- Should take advantage of additional structure