Training Selection for Tuning Entity Matching

Hanna Köpcke

Oberseminar Zingst 29.06.-03.07.2008
Outline

- Entity matching
- Entity Matching Framework
- Training data
- Selection strategies
- Evaluation
- Summary and outlook
Entity matching

The merge/purge problem for large databases - all 5 versions


Page 1. The Merge/Purge Problem for Large Databases * Abstract Mauricio A. Hernández Salvatore J. Stolfo {mrauricio, sal}@cs. columbia.edu ...
Cited by 342 - Related Articles - Web Search

2 [CITATION] The Merge/Purge Problem for Large Databases
Cited by 4 - Related Articles - Web Search

3 [CITATION] The merge/purge problem for large databases
MA Hernández SJ Stolfo - Proceedings of the ACM SIGMOD International Conference on ..., 1995
Cited by 4 - Related Articles - Web Search

[CITATION] The merge/purge problem for large datasets
MA Hernandez, SJ Stolfo - Proc. Of the SIGMOD, 1995
Cited by 2 - Related Articles - Web Search

[CITATION] Hernández, SJ Stolfo The merge/purge problem for large database
M Hern - Proceedings of the 1995 ACM SIGMOD International Conference ..., 1995
Cited by 3 - Related Articles - Web Search

1 Heterogeneous venue names
2 Extraction errors
3 Typos (author name)
4 Missing authors
Entity matching

- Given two sets of entities $A \subseteq S_A$ and $B \subseteq S_B$ of a particular semantic entity type from data sources $S_A$ and $S_B$, the entity matching (EM) problem is to identify all correspondences between entities in $A \times B$ representing the same real-world object.
Entity Matching Framework

Application

Blocking library

Blocking

Match candidates

Strategy application

Correspondences

C ⊆ S_A ∩ S_B

Specification

Training selection

Entity pairs

Training data

EM strategy generation

EM strategy

Selection methods library

Matcher Library

Learner Library

... ...

Threshold-Random

Jaccard

Support Vector Machine

Threshold-Equal

EditDistance

Decision Tree

SA

SB

Trigram Blocking

Sorted Neighbourhood

Matcher Library

... ...

Jaccard

Cosine

Logistic Regression

Training Selection for Tuning Entity Matching
Tuning approach

● Treat the objective of determining an EM strategy as a two-class (match or non-match) classification problem

● Employ supervised machine learning methods (learners)

● Requisite: training data
Training data

- Set of examples of matching and non-matching entity pairs

\[
\begin{pmatrix}
    x_{11}, & \ldots, & x_{1m}, y_1 \\
    \vdots & \ddots & \vdots \\
    x_{n1}, & \ldots, & x_{nm}, y_n \\
\end{pmatrix}
\]
Training data selection

- The effectiveness of a learner critically depends on the size and quality of the available training data.

- Requirements:
  - Representative for the entities to be matched
  - Exhibit the variety and distribution of errors observed in practice
  - Observation of differences between the available matcher algorithms so that an effective combination of different algorithms can be learned
  - Little manual overhead for labeling
Selection strategies

● Manual

● Semi-Automatic
  ▪ Random
  ▪ Threshold-Random
  ▪ Active Learning

● Automatic
  ▪ Nearest based
• **Threshold-Random** \((n, m, t)\): \(n\) object pairs are randomly selected among the ones satisfying a given minimal threshold \(t\) applying a similarity measure \(m\).

\[
\text{Trigram(title)} \geq 0.5
\]

\[
(x_1, \ldots, x_n, 1)
\]

\[
(x_1, \ldots, x_n, 0)
\]
Active Learning (1)

- Attempts to iteratively identify those pairs leading to maximal performance improvements when added to the training set
- Committee of n learners

**Training data**

\[
\begin{pmatrix}
  x_{11}, & \ldots, & x_{1m}, & y_1 \\
  \vdots & \ddots & \vdots & \vdots \\
  x_{n1}, & \ldots, & x_{nm}, & y_n
\end{pmatrix}
\]

**Learners**

\[
\begin{pmatrix}
  L_1 & L_2 & L_3 & L_4 & L_5 & L_6 & L_7 & L_8 & L_9 & L_{10} \\
  \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
  x_{11}, & \ldots, & x_{1m} \\
  \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
  x_{n1}, & \ldots, & x_{nm}
\end{pmatrix}
\]

**Labelling**

\[
\begin{pmatrix}
  L_1 & L_2 & L_3 & L_4 & L_5 & L_6 & L_7 & L_8 & L_9 & L_{10} \\
  1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
  0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
  0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 \\
  0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1
\end{pmatrix}
\]
Active Learning (2)

- Methods for creating committees:
  - Randomizing parameters
  - Partitioning training data
  - Attribute partition
Nearest based

- Proposed by Peter Christen et. al.*
- Selects entity pairs automatically, does not require manual labeling by a user
- The similarity vectors of the entity pairs are sorted according to their distances from the vectors containing only exact similarities and only total dissimilarities, respectively, and then selects the nearest entity pairs for training
- Distance measure: Manhattan distance

\[
d(x, y) = \sum_{i} |x_i - y_i|
\]

\[
\begin{bmatrix}
0.9, & 1.0, & 1.0, & 1.0, & 0.9 \\
0.0, & 0.0, & 0.0, & 0.0, & 0.0 \\
0.0, & 0.0, & 0.5, & 0.0, & 0.0 \\
0.7, & 0.3, & 0.5, & 0.7, & 0.9
\end{bmatrix}
\begin{bmatrix}
0.2 \\
5 \\
4.5 \\
1.9
\end{bmatrix}
\begin{bmatrix}
0.0, & 0.0, & 0.0, & 0.0, & 0.0 \\
0.0, & 0.0, & 0.5, & 0.0, & 0.0 \\
0.7, & 0.3, & 0.5, & 0.7, & 0.9
\end{bmatrix}
\begin{bmatrix}
4.8 \\
0 \\
0.5 \\
3.1
\end{bmatrix}
\]

Evaluation match tasks (1)

- Bibliographic domain
- Matching of publications

1. The merge/merge problem for large databases
   Mauricio A. Hernández, Salvatore J. Stolfo

   Many commercial organizations routinely gather large numbers of databases for various marketing and business analysis functions. The task is to correlate information from different databases by identifying distinct individuals that appear in a number ...

2. The merge/merge problem for large databases
   Mauricio A. Hernández, Salvatore J. Stolfo

   Many commercial organizations routinely gather large numbers of databases for various marketing and business analysis functions. The task is to correlate information from different databases by identifying distinct individuals that appear in a number ...
Evaluation match tasks (2)

- RIDDLE repository (http://www.cs.utexas.edu/users/ml/riddle/)
- Restaurant match problem
### Evaluation datasets

<table>
<thead>
<tr>
<th>match task</th>
<th># entities</th>
<th># attr.</th>
<th># corresp.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>source1</td>
<td>source2</td>
<td></td>
</tr>
<tr>
<td>Scholar-DBLP</td>
<td>64363</td>
<td>2616</td>
<td>4</td>
</tr>
<tr>
<td>ACM-DBLP</td>
<td>2294</td>
<td>2616</td>
<td>4</td>
</tr>
<tr>
<td>Restaurant</td>
<td>533</td>
<td>331</td>
<td>4</td>
</tr>
</tbody>
</table>
matcher configuration

- Trigram similarity
- Trigrams are character substrings of length three.
- Example: **zingst**
  
  **Trigrams:** \{##z, #zi, zin, ing, ngs, gst, st$, t$$\}
- String with length \(l \rightarrow l + 3 - 1\) Trigrams
- Similar strings have many common Trigrams

\[
\text{TrigramSim}(s_1, s_2) = \frac{2 \times |\text{Trigrams}(s_1) \cap \text{Trigrams}(s_2)|}{|\text{Trigrams}(s_1)| + |\text{Trigrams}(s_2)|}
\]
Support Vector Machine (SVM)

\[ h(\vec{x}) = \text{sign} \left( b + \sum_{i=1}^{n} w_i x_i \right) \]

\[ f_{\text{SVM}}(a,b) = \begin{cases} 
\text{match, if} & \left( 0.8 \cdot \text{Trigram}(title_a, title_b) + 0.8 \cdot \text{Trigram}(authors_a, authors_b) - 1.1 \right) > 0 \\
\text{non-match, otherwise} & 
\end{cases} \]
Evaluation of training selection strategies

- **Semi-automatic**
  - Random
  - Threshold-Random(n, TrigramSimilarity, 0.5)
  - Active Learning
    - 20 initial training examples
    - disjoint partitioning to train SVM 10 times
    - strategies for initial training selection:
      - Random
      - Threshold-Random
      - Nearest based

- **Automatic**
  - Nearest based
Evaluation measures

- Training time
- Application time
- Performance: F-Measure
Baseline strategies

- Bibliographic match tasks:
  - trigram similarity on both title and authors with a threshold of 0.5
  - Scholar-DBLP: 0.823 F-measure
  - ACM-DBLP: 0.914 F-measure

- Restaurant match task:
  - trigram similarity on name with threshold 0.8
  - 0.881 F-measure
Training time

Scholar-DBLP

ACM-DBLP

Restaurant

Training time (ms)

Number of training examples

Random
Threshold-Random
Nearest based
Random Active
Threshold-Random Active
Nearest based Active

Restaurant

Number of training examples

Training time (ms)
Application time

Scholar-DBLP

ACM-DBLP

Restaurant

application time (ms)

number of training examples

application time (ms)

number of training examples

application time (ms)

number of training examples

- Random
- Threshold-Random
- Nearest based
- Random Active
- Threshold-Random Active
- Nearest based Active
Performance

- **Scholar-DBLP**
  - Number of training examples: 10, 20, 30, 40, 50, 60, 70, 80, 90, 100
  - F-Measure:
    - Random
    - Threshold-Random
    - Nearest based
    - Random Active
    - Threshold-Random Active
    - Nearest based Active
    - Baseline

- **ACM-DBLP**
  - Number of training examples: 10, 20, 30, 40, 50, 60, 70, 80, 90, 100
  - F-Measure:
    - Random
    - Threshold-Random
    - Nearest based
    - Random Active
    - Threshold-Random Active
    - Nearest based Active
    - Baseline

- **Restaurant**
  - Number of training examples: 10, 20, 30, 40, 50, 60, 70, 80, 90, 100
  - F-Measure:
    - Random
    - Threshold-Random
    - Nearest based
    - Random Active
    - Threshold-Random Active
    - Nearest based Active
    - Baseline
Summary and outlook

- Training selection
- Evaluation
- Ongoing work
  - Time for training selection
  - Efficient implementation of similarity measures
  - Blocking
¡Thank You! ¿Questions?