A Generic and Flexible Framework for Selecting Correspondences in Matching and Alignment Problems

Fabien Duchateau

Université Claude Bernard Lyon 1 / LIRIS

DATA’2013 Conference, Reykjavík

http://liris.cnrs.fr/~fduchateau/
Large amount of data is produced everyday. For meaningful exploitation, this data has to be integrated:

- Fusioning catalogs of products
- Generating new knowledge from scientific databases
- Helping decision-makers during catastrophic scenarios

Discovering correspondences between data sources ⇒ schema matching, ontology alignment, entity resolution

Zohra Bellahsene, Angela Bonifati, and Erhard Rahm. 
*Schema Matching and Mapping.*
Springer-Verlag, Heidelberg, 2011.

Jérôme Euzenat and Pavel Shvaiko.  
*Ontology matching.*
Motivation Example

Two Web Forms about Hotel Booking

David Aumueller, Hong Hai Do, Sabine Massmann, and Erhard Rahm.
Schema and ontology matching with COMA++.
In ACM SIGMOD, pages 906–908, 2005.
Motivation Example

Discovering Correspondences for the Web forms with COMA++

David Aumueller, Hong Hai Do, Sabine Massmann, and Erhard Rahm.
Schema and ontology matching with COMA++.
In ACM SIGMOD, pages 906–908, 2005.
Outline of the Talk

Preliminaries

Details of the Framework
  A Model for Classifying Similarity Measures
  Detecting Discriminative Measures
  Computing a Confidence Score

Experimental Validation
  Experimental Protocol
  Experiment Results
Overview of the Matching/Alignment Problem
Overview of the Matching/Alignment Problem

Data sources

\{ a, b, \ldots \} → \{ a', b', \ldots \}

Candidate correspondences

\{ (a, a'), (a, b'), \ldots \}

Similarity measures

- Similarity measure 1
- Similarity measure 2
- Similarity measure N

Computation of similarity values

\{ \text{trigram}(a, a') = 0.45, \text{levenhstein}(a, a') = 0.3, \ldots \text{levenhstein}(k, n') = 0.1 \}
Overview of the Matching/Alignment Problem

- **Data Sources:** S1 \{ a, b, \ldots \}, S2 \{ a', b', \ldots \}
- **Candidate Correspondences:** \{ (a, a'), (a, b'), \ldots \}
- **Similarity Measures:**
  - Measure 1
  - Measure 2
  - Measure N
  \{ trigram, levenhstein, \ldots \}
- **Computation of Similarity Values**
  - Individual Scores:
    - \{ trigram(a, a') = 0.45, levenhstein(a, a') = 0.3, \ldots, levenhstein(k, n') = 0.1 \}
- **Global Scores:**
  - \{ (a, a') = 0.37, (a, b') = 0.12, \ldots, (k, n') = 0.03 \}
- **Combination of the Scores**
- **Weighted Average**
Overview of the Matching/Alignment Problem

- Data sources: \{ a, b, ... \}\{ a', b', ... \}
- Candidate correspondences: \{ (a, a'), (a, b'), ... \}
- Similarity measures:
  - Similarity measure 1
  - Similarity measure 2
  - Similarity measure N
  \{ trigram, levenshtein, ... \}
- Computation of Similarity Values
- Individual scores:
  - \{ trigram(a, a') = 0.45, levenshtein(a, a') = 0.3, ... levenshtein(k, n') = 0.1 \}
- Selection of Correspondences
- Global scores:
  - \{ (a, a') = 0.37, (a, b') = 0.12, ... (k, n') = 0.03 \}
- Combination of the Scores
- Weighted average
- Output correspondences: \{ (a, a'), (b, b'), ... \}
- Threshold
Issues

Tuning:

- Difficulty for tuning a similarity measure (e.g., weights, thresholds)
- Difficulty for tuning the combination function (e.g., strong impact of similarity measures of the same type)
- No extensibility (adding a new measure involves tuning again)

Selection of correspondences:

- All similarity values may not be significant for determining the relevance of a correspondence
- Inability of a similarity measure for discovering a correspondence (e.g., with two polysemous labels "mouse")
Proposition

A generic framework for selecting correspondences in matching/alignment problems:

▶ A classification of similarity measures according to their features

▶ Automatic selection of the meaningful similarity values to compute a confidence score

▶ No need for tuning

▶ Validation of the approach with a benchmark containing real-world entity matching datasets
Running Example

- Two data sources \(d\) and \(d'\):
  - \(E_d = \{a, b, c\}\)
  - \(E_{d'} = \{a', b', d'\}\)

- Set of correct correspondences: \(\{(a, a'), (b, b')\}\)

- Set of four similarity measures: \(\{\text{sim}_1, \text{sim}_2, \text{sim}_3, \text{sim}_4\}\)

<table>
<thead>
<tr>
<th>(\text{sim}_1)</th>
<th>a</th>
<th>b</th>
<th>c</th>
</tr>
</thead>
<tbody>
<tr>
<td>a'</td>
<td>0.8</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>b'</td>
<td>0</td>
<td>0.3</td>
<td>0</td>
</tr>
<tr>
<td>d'</td>
<td>0.8</td>
<td>0</td>
<td>0.7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(\text{sim}_2)</th>
<th>a</th>
<th>b</th>
<th>c</th>
</tr>
</thead>
<tbody>
<tr>
<td>a'</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>b'</td>
<td>0.2</td>
<td>0.1</td>
<td>0.2</td>
</tr>
<tr>
<td>d'</td>
<td>0.8</td>
<td>0.2</td>
<td>0.6</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(\text{sim}_3)</th>
<th>a</th>
<th>b</th>
<th>c</th>
</tr>
</thead>
<tbody>
<tr>
<td>a'</td>
<td>0.6</td>
<td>0.2</td>
<td>0.1</td>
</tr>
<tr>
<td>b'</td>
<td>0.3</td>
<td>0.9</td>
<td>0.4</td>
</tr>
<tr>
<td>d'</td>
<td>0.3</td>
<td>0.2</td>
<td>0.2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(\text{sim}_4)</th>
<th>a</th>
<th>b</th>
<th>c</th>
</tr>
</thead>
<tbody>
<tr>
<td>a'</td>
<td>0</td>
<td>0</td>
<td>0.5</td>
</tr>
<tr>
<td>b'</td>
<td>0</td>
<td>0.5</td>
<td>0</td>
</tr>
<tr>
<td>d'</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Similarity Matrices for Similarity Measures
Outline

Preliminaries

Details of the Framework
- A Model for Classifying Similarity Measures
- Detecting Discriminative Measures
- Computing a Confidence Score

Experimental Validation
- Experimental Protocol
- Experiment Results
**Intuition:** similarity measures can be organized according to various features, and a score can be computed to compare their ability for matching

- Category (e.g., terminological, linguistic, structural)
- Type of input (e.g., character strings, records)
- Type of output (e.g., number, semantic relationship)
- Use of external resources (e.g., a dictionary, an ontology)

---

**References**

A comparison of string distance metrics for name-matching tasks.

Pavel Shvaiko and Jerome Euzenat.
A survey of schema-based matching approaches.
*Journal of Data Semantics IV, pages 146–171, 2005.*
A Model for Classifying Similarity Measures (2)

Modelization of the similarity measures:

- Representation of a measure by a binary vector according to its features (1 for the feature, 0 else)

- Computation of a difference score $\Delta_{sim_i} \Rightarrow$ a similarity measure is different from the others if its vector is different. The more unique features a measure has, the more dissimilar it is w.r.t. other measures

- Computation of a dissimilarity score $\Rightarrow$ normalization of the difference score in $[0, 1]$

Result: each similarity measure obtains a dissimilarity score
Running Example

Binary Vectors for each Similarity Measure

<table>
<thead>
<tr>
<th>Feature</th>
<th>sim1</th>
<th>sim2</th>
<th>sim3</th>
<th>sim4</th>
</tr>
</thead>
<tbody>
<tr>
<td>terminology</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>structural</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>constraints</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>dictionary</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>ontology</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>element-level</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>relationship-level</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>semantic-result</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Difference and Dissimilarity Scores of each Measure

<table>
<thead>
<tr>
<th>Measure</th>
<th>sim1</th>
<th>sim2</th>
<th>sim3</th>
<th>sim4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Δ</td>
<td>0.33</td>
<td>0.33</td>
<td>0.67</td>
<td>0.375</td>
</tr>
<tr>
<td>dissim</td>
<td>0.19</td>
<td>0.19</td>
<td>0.40</td>
<td>0.22</td>
</tr>
</tbody>
</table>

The similarity measure \( sim_1 \) has 19% of different features compared to other measures, or \( sim_1 \) has an ignorance degree equal to 81%.
Detecting Discriminative Measures

**Intuition:** a matcher should identify the significant similarity values and the discriminative measures for a candidate correspondence

- For each similarity measure, use of the mean and the standard deviation to obtain a range of non-discriminative values
- A similarity value outside of that range and the associated measure are considered discriminative for a candidate correspondence
- One iteration may not be sufficient: discarding of the previous discriminative values for next iteration

Result: each candidate correspondence is associated to a set of discriminative similarity measures
Running Example

**Similarity Matrices for Similarity Measures**

- $\text{sim}_1$
  
<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>b</th>
<th>c</th>
</tr>
</thead>
<tbody>
<tr>
<td>a'</td>
<td>0.8</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>b'</td>
<td>0</td>
<td>0.3</td>
<td>0</td>
</tr>
<tr>
<td>d'</td>
<td>0.8</td>
<td>0</td>
<td>0.7</td>
</tr>
</tbody>
</table>

- $\text{sim}_2$
  
<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>b</th>
<th>c</th>
</tr>
</thead>
<tbody>
<tr>
<td>a'</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>b'</td>
<td>0.2</td>
<td>0.1</td>
<td>0.2</td>
</tr>
<tr>
<td>d'</td>
<td>0.8</td>
<td>0.2</td>
<td>0.6</td>
</tr>
</tbody>
</table>

- $\text{sim}_3$
  
<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>b</th>
<th>c</th>
</tr>
</thead>
<tbody>
<tr>
<td>a'</td>
<td>0.6</td>
<td>0.2</td>
<td>0.1</td>
</tr>
<tr>
<td>b'</td>
<td>0.3</td>
<td>0.9</td>
<td>0.4</td>
</tr>
<tr>
<td>d'</td>
<td>0.3</td>
<td>0.2</td>
<td>0.2</td>
</tr>
</tbody>
</table>

- $\text{sim}_4$
  
<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>b</th>
<th>c</th>
</tr>
</thead>
<tbody>
<tr>
<td>a'</td>
<td>0</td>
<td>0</td>
<td>0.5</td>
</tr>
<tr>
<td>b'</td>
<td>0</td>
<td>0.5</td>
<td>0</td>
</tr>
<tr>
<td>d'</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

- $\text{Avg}_{\text{sim}_1} = 0.28$
- $\text{Std}_{\text{sim}_1} = 0.35$
- Range of non-discriminative values for $\text{sim}_1 = [0, 0.63]$
- Discriminative measures for $(a, a') = \{\text{sim}_1, \text{sim}_3\}$

1 All underlined values in the similarity matrices indicate that the measure is discriminative for the candidate correspondence at iteration 1.
Computing a Confidence Score (1)

**Intuition:** a confidence score should be higher for a candidate correspondence which obtains discriminative values with different similarity measures

- The confidence score is computed with the discriminative values and the dissimilarity scores

\[ \text{conf}^t_{(e,e')} = \sum_{i=1}^{n} \text{dissim}_i \times \frac{\sum_{i=1}^{n} \text{sim}_i(e,e')}{n} \]

- Solve conflict by discarding correspondences with already matched elements, or use refine technique to detect a complex correspondance

**Result:** each candidate correspondence obtains a confidence score
### Running Example

<table>
<thead>
<tr>
<th></th>
<th>sim1</th>
<th>sim2</th>
<th>sim3</th>
<th>sim4</th>
</tr>
</thead>
<tbody>
<tr>
<td>a’</td>
<td>0.8</td>
<td>0.1</td>
<td>0.6</td>
<td>0</td>
</tr>
<tr>
<td>b’</td>
<td>0</td>
<td>0.2</td>
<td>0.3</td>
<td>0</td>
</tr>
<tr>
<td>c’</td>
<td>0</td>
<td>0.3</td>
<td>0.2</td>
<td>0.5</td>
</tr>
<tr>
<td>d’</td>
<td>0.8</td>
<td>0.8</td>
<td>0.3</td>
<td>0</td>
</tr>
</tbody>
</table>

### Similarity Matrices for Similarity Measures

1. \( \text{conf}(b, b') = 0.43 \)
2. \( \text{conf}(a, a') = 0.41 \)
3. \( \text{conf}(a, d') = 0.30 \)
4. \( \text{conf}(c, d') = 0.25 \)
5. \( \text{conf}(c, a') = 0.19 \)
Running Example

Similarity Matrices for Similarity Measures

1. \( \text{conf}(b, b') = 0.43 \)
2. \( \text{conf}(a, a') = 0.41 \)
3. \( \text{conf}(a, d') = 0.30 \) discarded
4. \( \text{conf}(c, d') = 0.25 \) requires manual verification
5. \( \text{conf}(c, a') = 0.19 \) discarded
Outline

Preliminaries

Details of the Framework
   A Model for Classifying Similarity Measures
   Detecting Discriminative Measures
   Computing a Confidence Score

Experimental Validation
   Experimental Protocol
   Experiment Results
Experimental Protocol (1)

Benchmark for entity resolution

- Domains: Web products (Abt/Buy and Amazon/GoogleProducts) and publications (DBLP/Scholar and DBLP/ACM)
- Sizes: from 1081 entities (Abt) to 65000 (Scholar)
- Set of perfect correspondences: from 1097 (Abt-Buy) to 5347 (DBLP-Scholar)
- Tested with a matching tool: BenchTool

Hanna Kopcke, Andreas Thor, and Erhard Rahm.
Learning-based approaches for matching web data entities.
Experimental Protocol (2)

Our framework has been implemented:
- Use of 10 similarity measures (Second String API\textsuperscript{2}, Resnik metric with Wordnet, a contextual measure)
- Classification of the measures with 8 features

What we demonstrate?
- Robustness and extensibility
- Matching quality at least equal to BenchTool

\begin{itemize}
  \item \textbf{Fabien Duchateau, Remi Coletta, Zohra Bellahsene, and Renée J. Miller.}
  \textit{(Not) Yet Another Matcher.}
  \textit{In Conference on Information and Knowledge Management, pages 1537–1540, 2009.}
  \item \textbf{Philip Resnik.}
  \textit{Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language.}
  \textit{Journal of Artificial Intelligence Research, 11:95–130, 1999.}
\end{itemize}

\textsuperscript{2}http://secondstring.sourceforge.net/
Demonstrating Robustness and Extensibility

Quality results according to the number of similarity measures:

- Random selection of the measures, average results of 10 runs
- Without any tuning, our approach integrates new measures
- The matching quality increases with more available measures
Comparative results in terms of F-measure:

- Web products are more difficult to match: confusing attribute "description" (full sentences) and some very similar products (e.g., HD with different storage capacity)
- Our approach improves over Benchtool for the four datasets
Conclusion

Contributions:

▶ A generic and extensible framework for selecting correspondences, with no need for tuning
▶ Validation of the approach with an entity matching benchmark

Perspectives:

▶ More experiments (with schemas/ontologies/parameters)
▶ Study the replacement of boolean vectors by real vectors
▶ Automatically determine the features of a similarity measure, using a benchmark (e.g., OAEI benchmark track) or the value distribution of the measure

Ontology Alignment Evaluation Initiative (OAEI).
Thank you!

Questions are guaranteed in life; Answers aren't.