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/~fduchate/

Context

Large amount of data is produced everyday. For meaningful exploitation, this data has to be integrated:

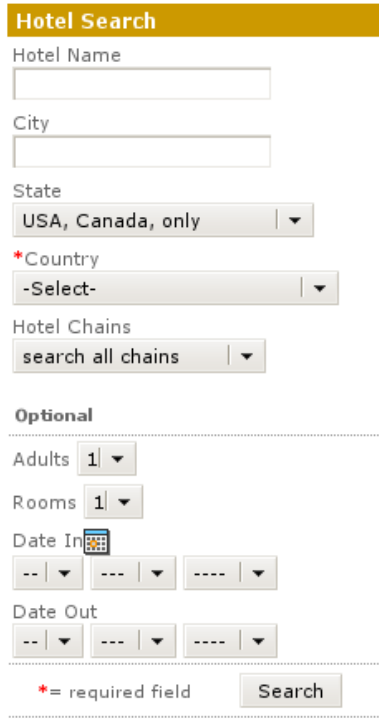
- Fusioning catalogs of products
- Generating new knowledge from scientific databases
- Helping decision-makers during catastrophic scenarios

Discovering correspondences between data sources $\Rightarrow$ 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.
Springer-Verlag, Heidelberg (DE), 2007.

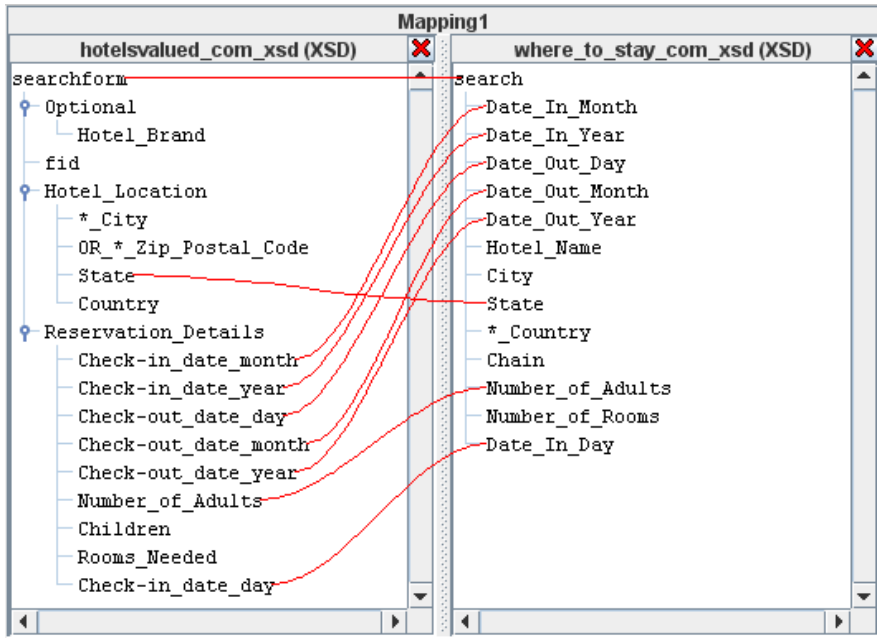
Motivation Example

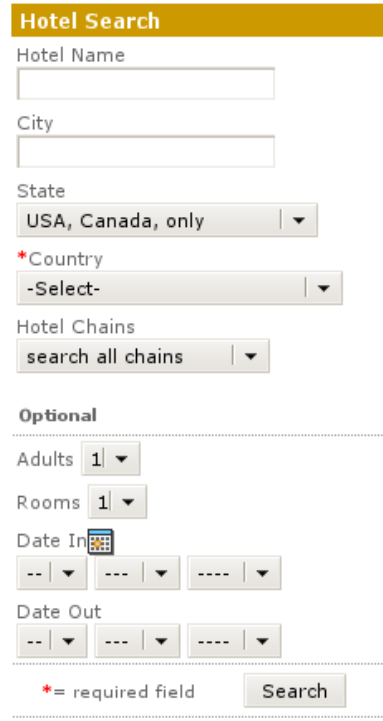


Two Web Forms about Hotel Booking

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



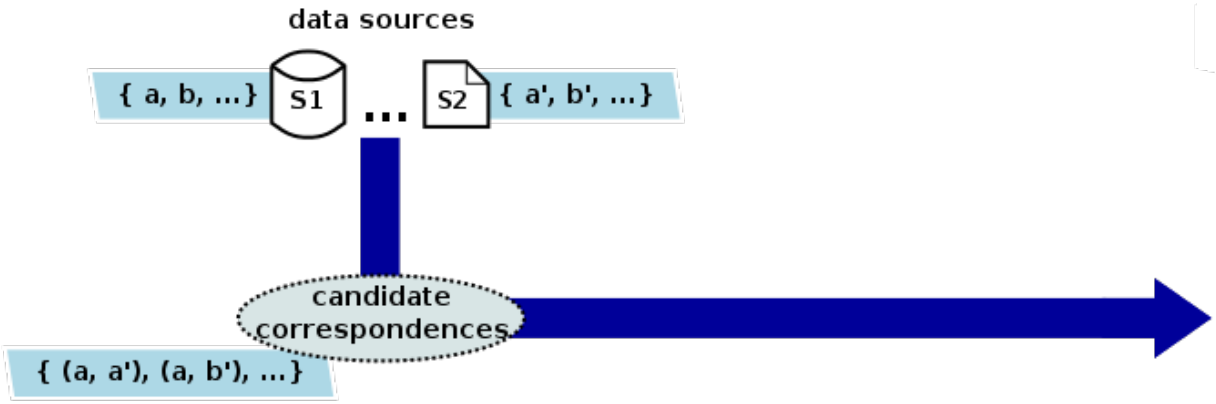


Discovering Correspondences for the Web forms with COMA++

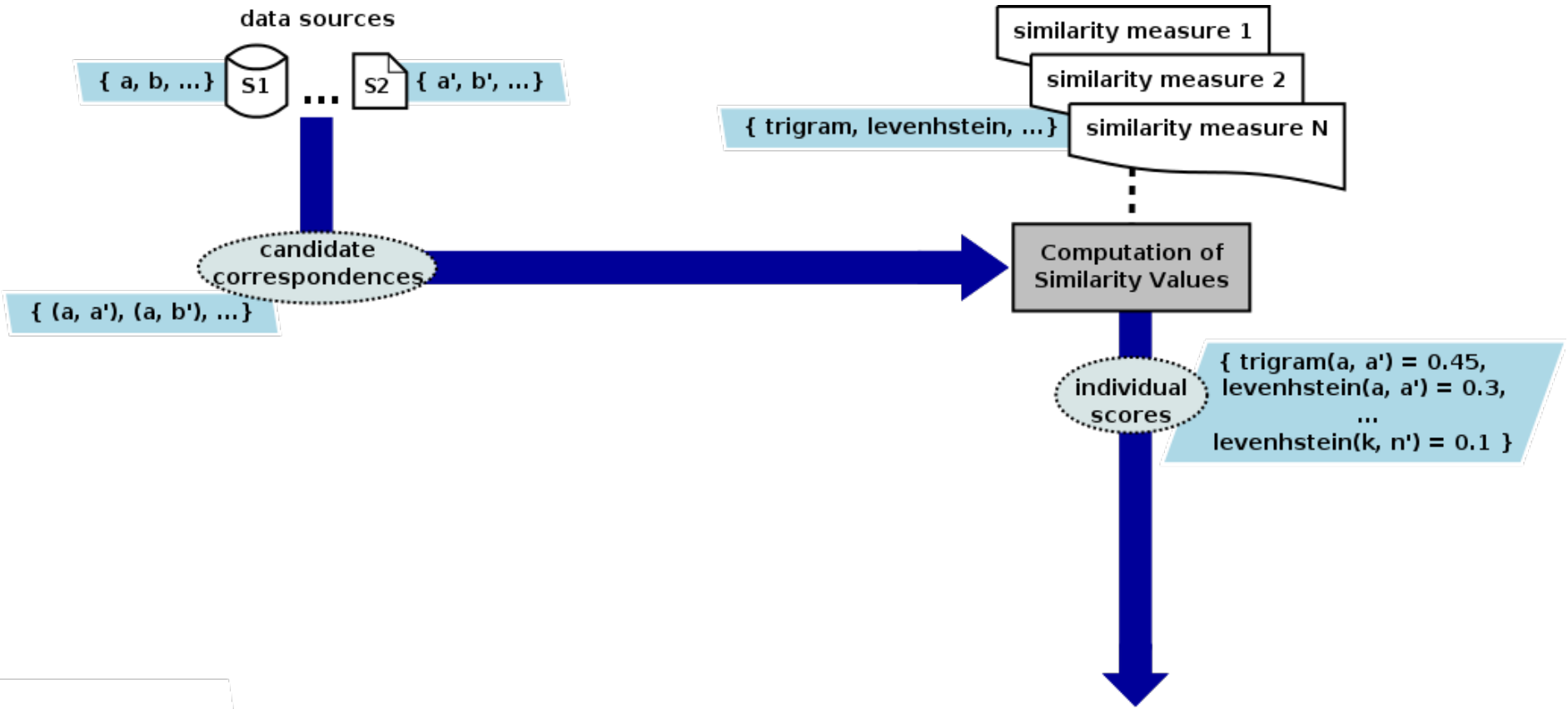
David Aumueller, Hong Hai Do, Sabine Massmann, and Erhard Rahm.
Schema and ontology matching with COMA++.
In ACM SIGMOD, pages 906–908, 2005.

Outline of the Talk

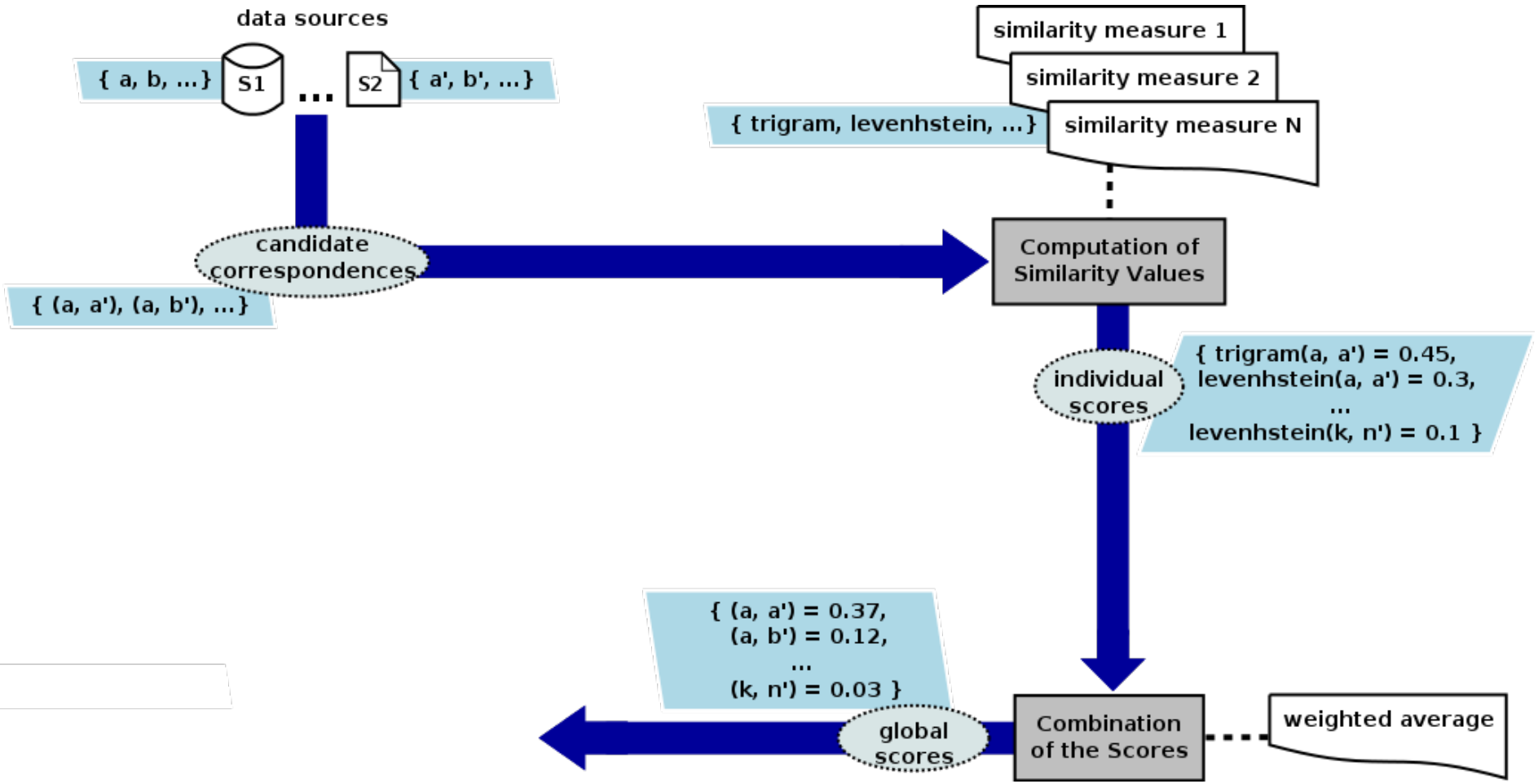
Overview of the Matching/Alignment Problem



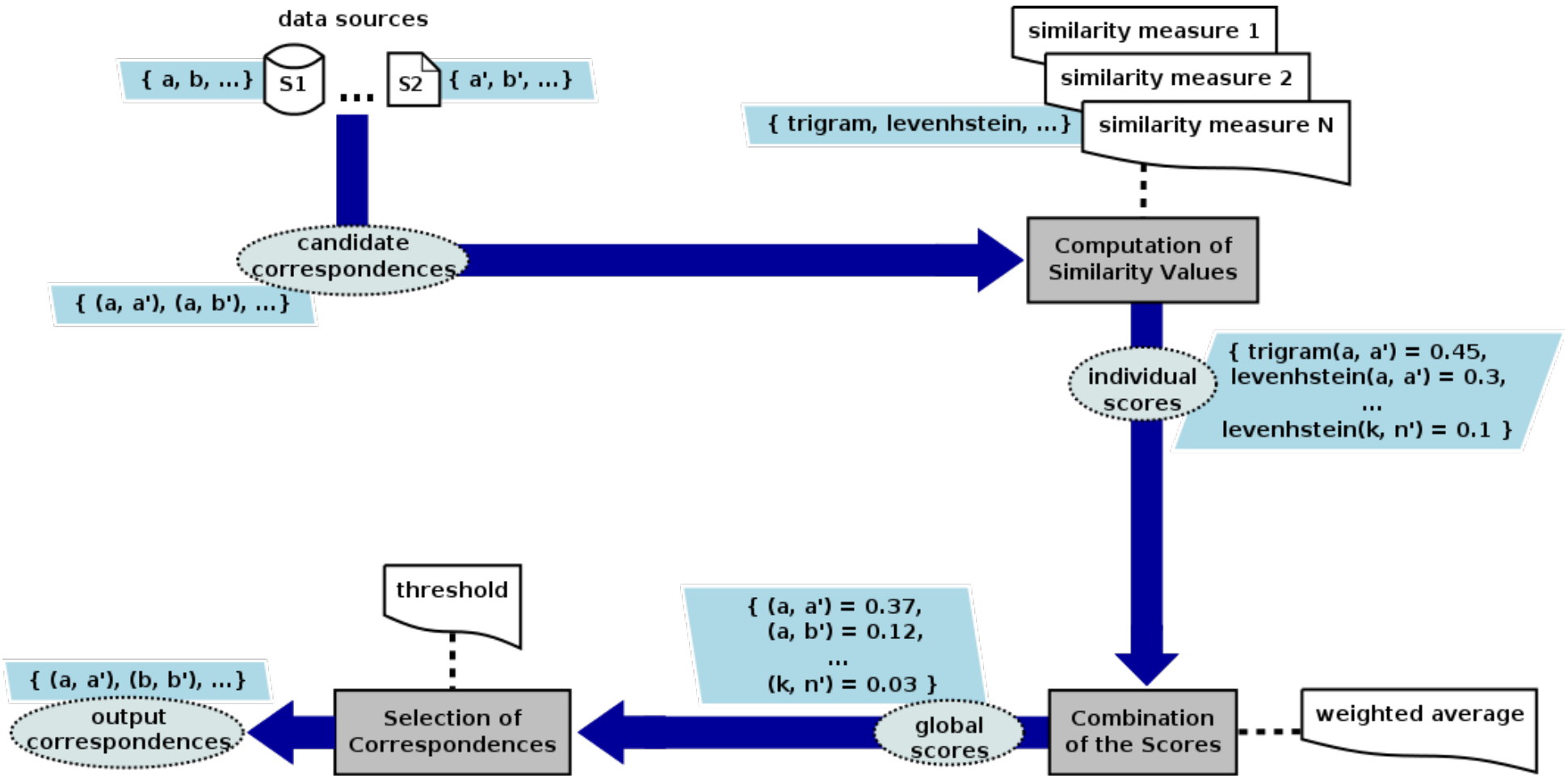
Overview of the Matching/Alignment Problem



Overview of the Matching/Alignment Problem



Overview of the Matching/Alignment Problem



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':
  - $\mathcal{E}_d = \{a, b, c\}$
  - $\mathcal{E}_{d'} = \{a', b', d'\}$
- Set of correct correspondences: $\{(a, a'), (b, b')\}$
- Set of four similarity measures: $\{sim_1, sim_2, sim_3, sim_4\}$

sim_1	a	b	c
a'	0.8	0	0
b'	0	0.3	0
d'	0.8	0	0.7

sim_2	a	b	c
a'	0.1	0.1	0.1
b'	0.2	0.1	0.2
d'	0.8	0.2	0.6

sim_3	a	b	c
a'	0.6	0.2	0.1
b'	0.3	0.9	0.4
d'	0.3	0.2	0.2

sim_4	a	b	c
a'	0	0	0.5
b'	0	0.5	0
d'	0	0	0

Similarity Matrices for Similarity Measures

Outline

A Model for Classifying Similarity Measures (1)

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)

W. Cohen, P. Ravikumar, and S. Fienberg.
A comparison of string distance metrics for name-matching tasks.
In Proceedings of the IJCAI, 2003.

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

	terminological	structural	constraints	dictionary	ontology	element-level	relationship-level	semantic-result
sim1	1	0	0	0	0	1	0	0
sim2	1	0	0	0	0	1	0	0
sim3	0	1	1	1	0	1	1	0
sim4	0	0	0	0	1	1	0	1

Binary Vectors for each Similarity Measure

	sim_1	sim_2	sim_3	sim_4
Δ	0.33	0.33	0.67	0.375
dissim	0.19	0.19	0.40	0.22

Difference and Dissimilarity Scores of each Measure

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

sim_1	a	b	c
a'	0.8	0	0
b'	0	0.3	0
d'	0.8	0	0.7

sim_2	a	b	c
a'	0.1	0.1	0.1
b'	0.2	0.1	0.2
d'	0.8	0.2	0.6

sim_3	a	b	c
a'	0.6	0.2	0.1
b'	0.3	0.9	0.4
d'	0.3	0.2	0.2

sim_4	a	b	c
a'	0	0	0.5
b'	0	0.5	0
d'	0	0	0

Similarity Matrices for Similarity Measures[1]

- $\text{Avg}_{sim_1} = 0.28$
- $\text{Std}_{sim_1} = 0.35$
- Range of non-discriminative values for $\text{sim}_1 = [0, 0.63]$
- Discriminative measures for (a, a') $= \{sim_1, sim_3\}$

[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

$$conf^{t}_{(e,e')} = \sum_{i=1}^{n} dissim_{sim_i} \times \frac{\sum_{i=1}^{n} sim_i(e, e')}{n}$$

- 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

sim_1	a	b	c
a'	0.8	0	0
b'	0	0.3	0
d'	0.8	0	0.7

sim_2	a	b	c
a'	0.1	0.1	0.1
b'	0.2	0.1	0.2
d'	0.8	0.2	0.6

sim_3	a	b	c
a'	0.6	0.2	0.1
b'	0.3	0.9	0.4
d'	0.3	0.2	0.2

sim_4	a	b	c
a'	0	0	0.5
b'	0	0.5	0
d'	0	0	0

Similarity Matrices for Similarity Measures

1. conf(b, b') = 0.43
2. conf(a, a') = 0.41
3. conf(a, d') = 0.30
4. conf(c, d') = 0.25
5. conf(c, a') = 0.19

Running Example

sim_1	a	b	c
a'	0.8	0	0
b'	0	0.3	0
d'	0.8	0	0.7

sim_2	a	b	c
a'	0.1	0.1	0.1
b'	0.2	0.1	0.2
d'	0.8	0.2	0.6

sim_3	a	b	c
a'	0.6	0.2	0.1
b'	0.3	0.9	0.4
d'	0.3	0.2	0.2

sim_4	a	b	c
a'	0	0	0.5
b'	0	0.5	0
d'	0	0	0

Similarity Matrices for Similarity Measures

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

Outline

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.
IEEE Internet Computing, 14(4):23–31, 2010.

Experimental Protocol (2)

Our framework has been implemented:

- Use of 10 similarity measures (Second String API[2], Resnik metric with Wordnet, a contextual measure)
- Classification of the measures with 8 features

What we demonstrate ?

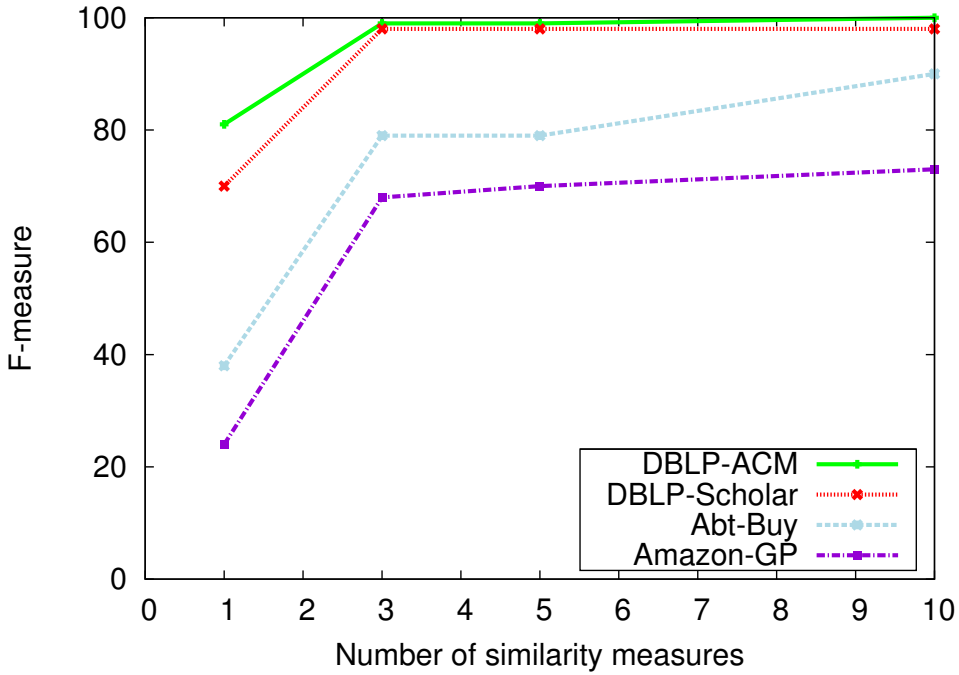
- Robustness and extensibility
- Matching quality at least equal to BenchTool

Fabien Duchateau, Remi Coletta, Zohra Bellahsene, and Renée J. Miller.
(Not) Yet Another Matcher.
In Conference on Information and Knowledge Management, pages 1537–1540, 2009.

Philip Resnik.
Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language.
Journal of Artificial Intelligence Research, 11:95–130, 1999.

[2] http://secondstring.sourceforge.net/

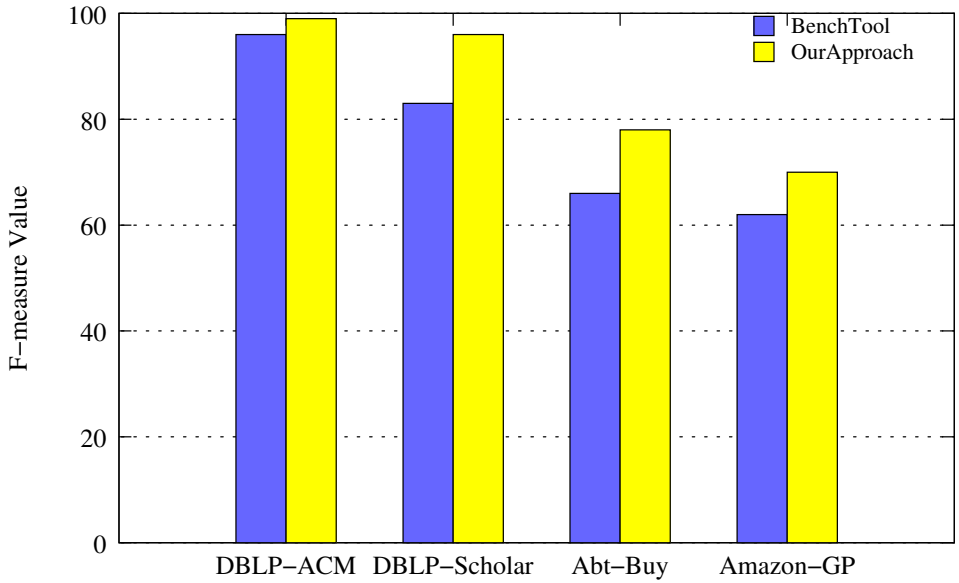
Demonstrating Robustness and Extensibility



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

Demonstrating Matching Quality



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 :

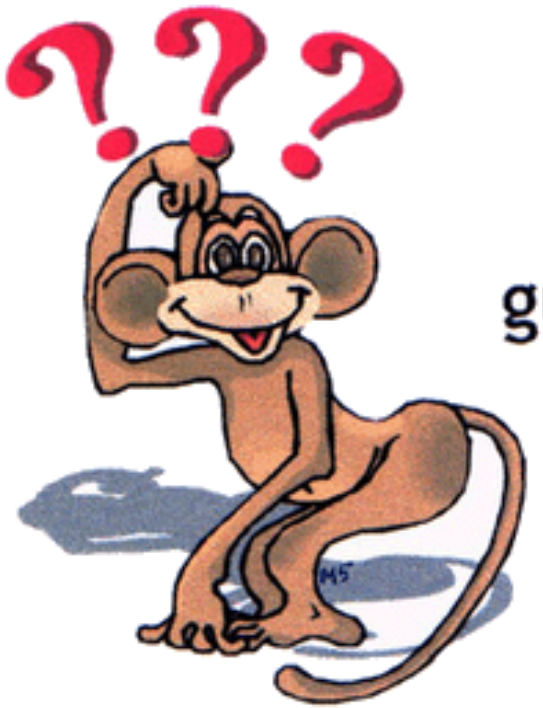
- A generic and extensible framework for selecting correspondences, with no need for tuning
- Validation of the approach with an entity matching benchmark

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).
http://oaei.ontologymatching.org/, 2013.

Thank you !



Questions are guaranteed in life; Answers aren't.