Evaluating Data and Instance Matching

The Instance Matching Track at OAEI 2009

Alfio Ferrara[1] Andriy Nikolov[2] François Scharffe[3]

[1]Università degli Studi di Milano
ferrara@dico.unimi.it :: http://islab.dico.unimi.it/ferrara

[2]The Open University

[3]INRIA

October 25 2009, Chantilly USA

Motivations and goals

- In the semantic web, data are often represented as ontology instances and/or RDF graphs.
- For both theoretical and practical reasons, it is important to have reliable techniques and tools for automatically comparing different data descriptions in order to find instances referred to the same real world objects.
- Several tools for ontology matching provide also instance matching functionalities.
- Thus, we decided this year to organize this 1st edition of the Instance Matching track at OAEI 2009.

Some differences between ontology and instance matching

- Instance matching requires usually to match very large datasets.
- Linguistic matching techniques, dictionaries, thesauri are usually less useful for data such as person names, ages, emails, etc.
- Useful information for matching purposes is available both in the instance descriptions and in the schema/ontology.
- Finally, any instance can describe more than one object, depending on the context and the matching goals.

How we organized the track

- The instance matching track has been organized in two sub-tracks:
- AKT-Rexa-DBLP
  - Focused on real datasets
  - Expected mappings manually defined
  - Large datasets
- IIMB
  - Automatically generated from a real dataset
  - Expected mappings obtained by introducing controlled modifications
  - Small datasets

Participants

Track/System	**AFlood**	**ASMOV**	**DSSim**	**HMatch**	**FBEM**	**RiMOM**
AKT-Rexa-DBLP		(✓)	✓	✓	✓	✓
IIMB	✓	✓	✓	✓	✓	✓

Description of AKT-Rexa-DBLP

The AKT-Rexa-DBLP track aims at testing the capability of the tools to match individuals over three datasets. All three datasets were structured using the same schema.

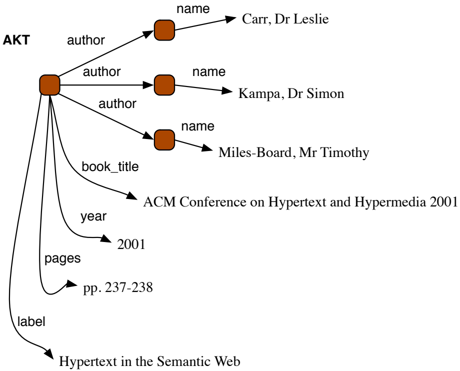
- AKT EPrints archive (http://eprints.aktors.org)
  - Information about papers produced within the AKT research project
  - About 847 instances, 2700 assertions
- Rexa (http://www.rexa.info)
  - Extracted from the Rexa search server, which was constructed at the University of Massachusetts using automatic information extraction algorithms
  - About 14.700 instances, more than 165.000 assertions
- SWETO DBLP (http://lsdis.cs.uga.edu/projects/semdis/swetodblp/)
  - Publicly available dataset listing publications from the computer science domain
  - About 1.500.000 instances, more than 2.000.000 assertions

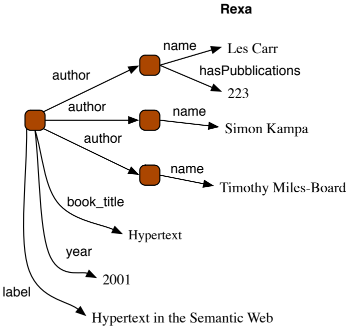
Challenges

- Authors are represented as instances of the foaf:Person class, and a special class sweto:Publication is defined for publications

- The challenges for the matchers included ambiguous labels (person names and paper titles) and noisy data (some sources contained incorrect information)

Examples

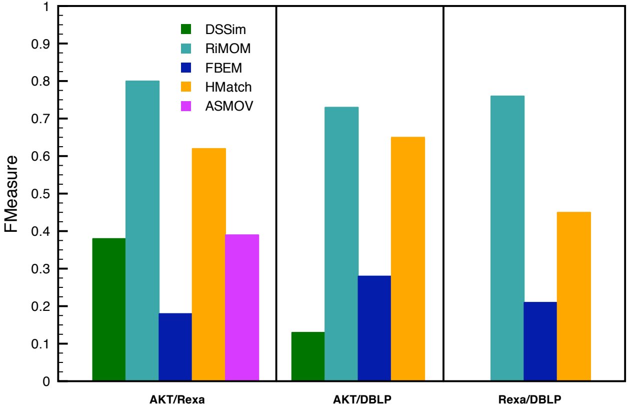

AKT
author
name
Carr, Dr Leslie
author
name
Kampa, Dr Simon
author
name
Miles-Board, Mr Timothy
book_title
ACM Conference on Hypertext and Hypermedia 2001
year
2001
pages
pp. 237-238
label
Hypertext in the Semantic Web


Rexa
name
Les Carr
hasPubblications
223
author
author
name
Simon Kampa
author
name
Timothy Miles-Board
book_title
Hypertext
year
2001
label
Hypertext in the Semantic Web

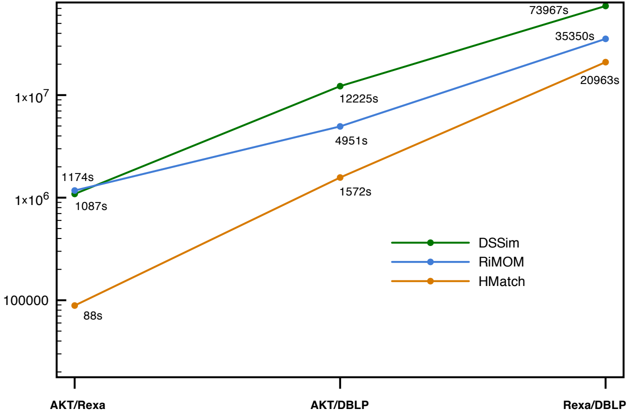
Results

	sweto:Publication			foaf:Person			Overall		
System	Prec.	Rec.	FMeas.	Prec.	Rec.	FMeas.	Prec.	Rec.	FMeas.
AKT/Rexa									
DSSim	0.15	0.16	0.16	0.81	0.30	0.43	0.60	0.28	0.38
RiMOM	1.00	0.72	0.84	0.92	0.70	0.79	0.93	0.70	0.80
FBEM	0.99	0.61	0.76	0.73	0.02	0.03	0.94	0.10	0.18
HMatch	0.97	0.89	0.93	0.94	0.39	0.56	0.95	0.46	0.62
ASMOV	0.32	0.79	0.46	0.76	0.24	0.37	0.52	0.32	0.39
AKT/DBLP									
DSSim	0	0	0	0.15	0.19	0.17	0.11	0.15	0.13
RiMOM	0.96	0.97	0.96	0.93	0.50	0.65	0.94	0.59	0.73
FBEM	0.98	0.80	0.88	0	0	0	0.98	0.16	0.28
HMatch	0.93	0.97	0.95	0.58	0.57	0.57	0.65	0.65	0.65
Rexa/DBLP									
DSSim	0	0	0	0	0	0	0	0	0
RiMOM	0.94	0.95	0.94	0.76	0.66	0.71	0.80	0.72	0.76
FBEM	0.98	0.15	0.26	1.00	0.11	0.20	0.99	0.12	0.21
HMatch	0.45	0.96	0.61	0.40	0.34	0.37	0.42	0.48	0.45

Comparison


1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
FMeasure
🟩 DSSim
🟦 RiMOM
🟦 FBEM
🟧 HMatch
🟪 ASMOV
AKT/Rexa
AKT/DBLP
Rexa/DBLP

Time performances



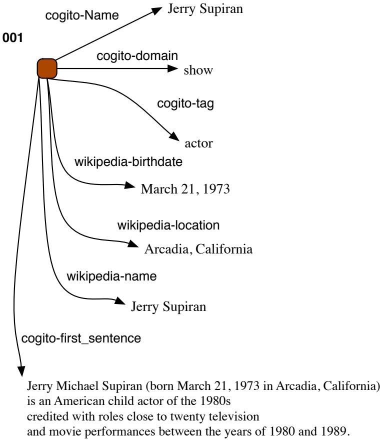
Description of IIMB

- The ISLab Instance Matching Benchmark (IIMB) is a benchmark automatically generated starting from one data source that is automatically modified according to various criteria
- The original data source contains OWL/RDF data about actors, sport persons, and business firms provided by the OKKAM European project (http://www.okkam.org)
- The benchmark is composed by 37 test cases. For each test case we require participants to match the original data source against a new data source. The original data source contains about 200 different instances. Each test case contains a modified version of the original data source and the corresponding reference alignment containing the expected results

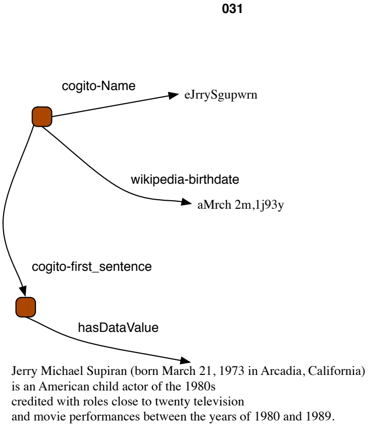
Challenges

- Test case 001: Contains an identical copy of the original data source (instance IDs are randomly changed)
- Test case 002 - Test case 010: Value transformations (i.e., typographical errors simulation, use of different standard for representing the same information)
- Test case 011 - Test case 019: Structural transformations (i.e., deletion of one or more values, transformation of datatype properties into object properties, separation of a single property into more properties)
- Test case 020 - Test case 029: Logical transformations (i.e., instantiation of identical individuals into different subclasses of the same class, instantiation of identical individuals into disjoint classes, instantiation of identical individuals into different classes of an explicitly declared class hierarchy)
- Test case 030 - Test case 037: Several combinations of the previous transformations

Examples



Jerry Michael Supiran (born March 21, 1973 in Arcadia, California)
is an American child actor of the 1980s
credited with roles close to twenty television
and movie performances between the years of 1980 and 1989.



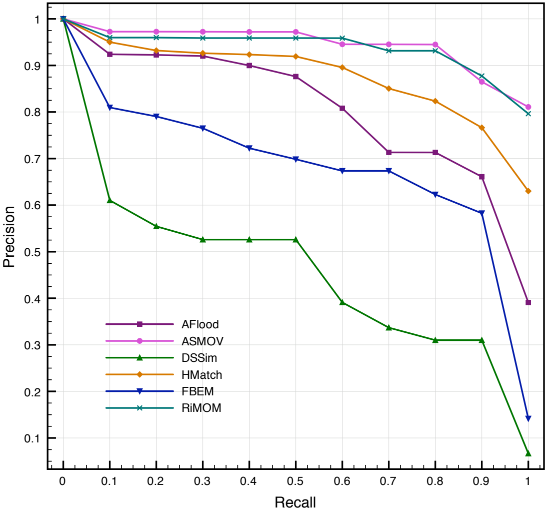
Jerry Michael Supiran (born March 21, 1973 in Arcadia, California)
is an American child actor of the 1980s
credited with roles close to twenty television
and movie performances between the years of 1980 and 1989.

Results

System	AFlood			ASMOV			DSSim		
Test	Prec.	Rec.	FMeas.	Prec.	Rec.	FMeas.	Prec.	Rec.	FMeas.
002 - 010	1.00	0.99	0.99	1.00	1.00	1.00	1.00	0.37	0.54
011 - 019	0.90	0.72	0.80	0.99	0.92	0.96	0.99	0.28	0.43
020 - 029	0.85	1.00	0.92	1.00	1.00	1.00	0.85	0.99	0.91
030 - 037	0.94	0.75	0.83	1.00	0.98	0.99	1.00	0.30	0.46
H-means	0.92	0.87	0.89	1.00	0.98	0.99	0.92	0.48	0.63

System	HMatch			FBEM			RiMOM		
Test	Prec.	Rec.	FMeas.	Prec.	Rec.	FMeas.	Prec.	Rec.	FMeas.
002 - 010	0.97	0.98	0.97	0.95	0.93	0.94	1.00	1.00	1.00
011 - 019	0.88	0.83	0.85	0.78	0.52	0.62	1.00	0.93	0.97
020 - 029	0.78	1.00	0.88	0.08	1.00	0.15	0.85	1.00	0.92
030 - 037	0.94	0.89	0.92	0.10	0.53	0.16	1.00	0.99	0.99
H-means	0.89	0.93	0.91	0.16	0.75	0.27	0.96	0.98	0.97

Comparison


Precision
Recall
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
AFlood
ASMOV
DSSim
HMatch
FBEM
RiMOM

Conclusion

- Good experience with quite good participation. Something to fix, but the session can be repeated
- Still something to do in instance matching and much to do in evaluating instance matching
- More emphasis on performances and quality. We need more complex data and schemas.

Thanks

- Results (continuously updated) are online at http://islab.dico.unimi.it/content/oaei2009/
- Comments, suggestions, ideas are more than welcome at ferrara@dico.unimi.it