INTERLINKING RDF-BASED DATASETS:

A STRUCTURE-BASED APPROACH

Pierre-Henri Paris, Fayçal Hamdi, Samira Si-Said Cherfi

INSTANCE MATCHING

WHAT?

- Linking instances together
- Specifying identical instances (owl:sameAs links)
- Interlink datasets

INSTANCE MATCHING

WHAT?

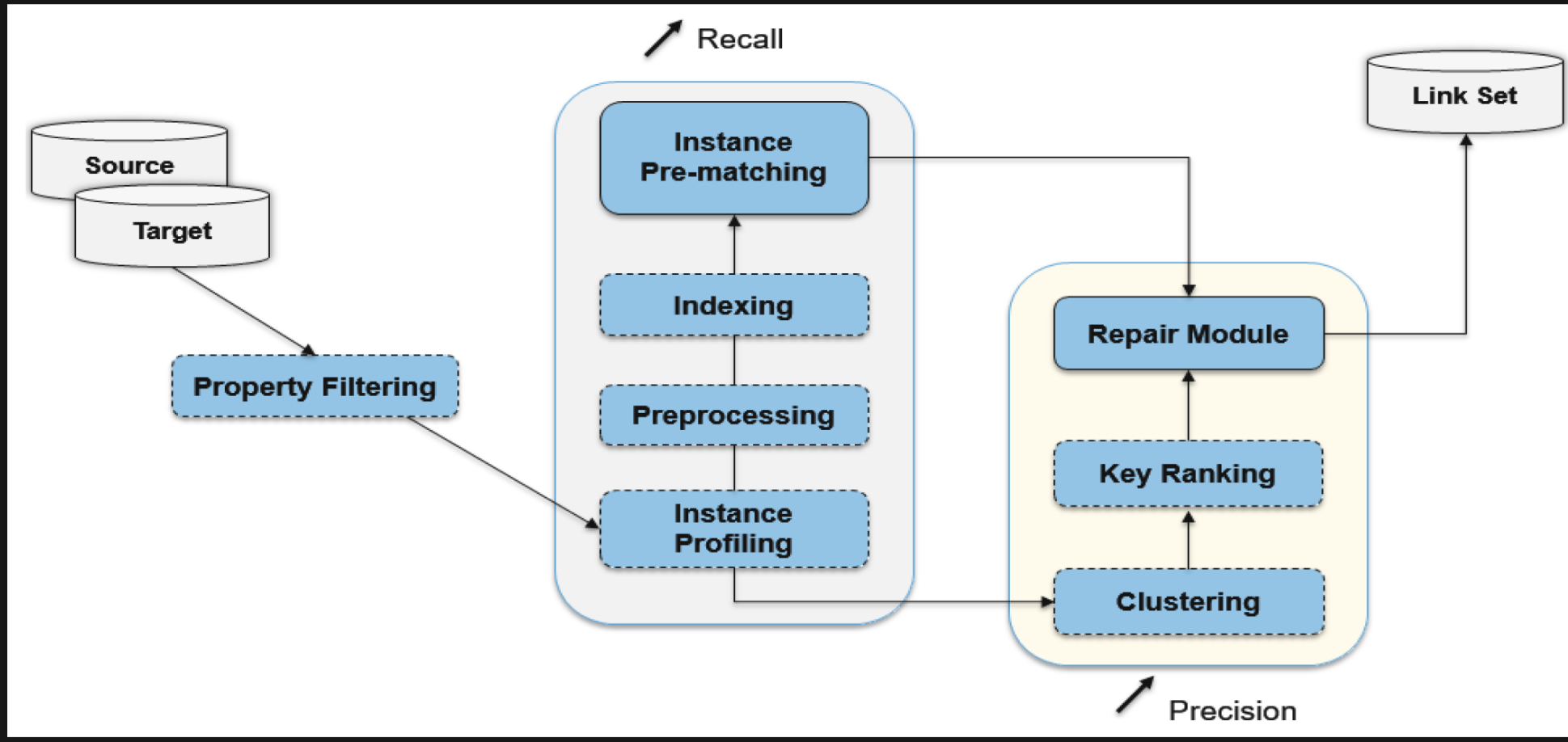
- Linking instances together
- Specifying identical instances (owl:sameAs links)
- Interlink datasets

WHY?

- Discover new knowledge (indiscernibility of identicals)
- Data integration
- etc.

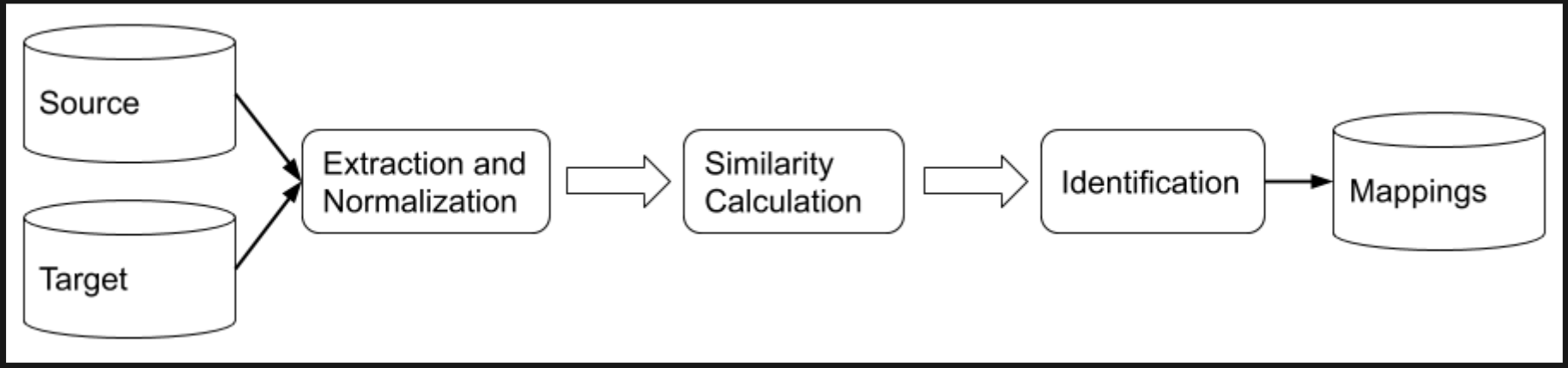
RELATED WORK

LEGATO



RELATED WORK

I-MATCH



- Ferraram, A., Nikolov, A., Scharffe, F., 2013. Data linking for the semantic web. Semantic Web: Ontology and Knowledge Base Enabled Tools, Services, and Applications 169, 326.
- Achichi, M., Bellahsene, Z., Todorov, K., 2016. A survey on web data linking. Revue des Sciences et Technologies de l'Information-Série ISI:Ingénierie des Systèmes d'Information.
- Nentwig, M., Hartung, M., Ngonga Ngomo, A.C., Rahm, E., 2017. A survey of current link discovery frameworks. Semantic Web 8, 419–436.
- External KBs, NLP, network measures, semantics, data mining, etc.

APPROACH

- Direct semantic proof
- The use of properties

APPROACH

DIRECT SEMANTIC PROOF

- Functional properties
- Maximum cardinality of properties
- etc.

APPROACH

DIRECT SEMANTIC PROOF

- Functional properties
- Maximum cardinality of properties
- etc.

EXAMPLE

If *hasFather is a functional property, (:John, hasFather, ns1:Bill) and (:John, hasFather, ns2:William)*
then *(ns1:Bill, owl:sameAs, ns2:William)*

APPROACH

THE USE OF PROPERTIES

First intuition (weight of a role):

If 90% of the People's instances use the role *name* but only 8% of those instances use the role *ownerOf*, then *ownerOf* might help more to determine (the absence of) an identity relation between two instances.

APPROACH

THE USE OF PROPERTIES

Second intuition (discriminating power of a role-value pair):

If we have 100 instances with the role-value <town, Paris> but only 3 instances with the role-value <town, Peyrabout>, then the couple <town, Peyrabout> helps to discriminate more instances.

APPROACH

THE USE OF PROPERTIES

Weight of clue:

If x1, x2 and x3 are three instances where x1 is from the source KB and x2 and x3 are from the target KB. If we have four clues between x1 and x2, and eight clues between x1 and x3 then we give a bonus to the comparison with the more clues to present.

APPROACH

THE USE OF PROPERTIES

Depth of a concept:

If $\mathcal{KB} = dbo$ then $depth_{dbo}(Agent) = 1$ and $depth_{dbo}(Biologist) = 4$ since $Agent$ is a direct sub concept of $owl{:}Thing$ and $Biologist \sqsubseteq Scientist \sqsubseteq Person \sqsubseteq Agent \sqsubseteq owl$

APPROACH

MAIN ALGORITHM

```
if IsSemProof(x1, x2):
  return SemProofValue(x1, x2)
scores = []
C = deepest common concept between x1 and x2
for R in {common roles between x1 and x2}:
  (maxSim, o) = max(R, x1, x2)
  subscore = Aggregation_1(
     maxSim,
     (1 - WKBs(R, C)),
     (1 - DKBs(C, R, o)))
  scores.append(subscore)
return Aggregation_2(weight of clue, scores)
```

EXPERIMENTS

DBPEDIA AND WIKIDATA

Our **goal** is to evaluate our approach on real-world datasets.

EXPERIMENTS

DBPEDIA AND WIKIDATA

Construction of source and target KBs

- DBpedia 2016-10 and DBpedia-Wikidata 03.30.2015
- From DBpedia, selection of 36 people each having at least 15 homonyms in Wikidata (rdfs:label)
- Source KB contains all statements having one of this 36 people in subject or object
- Target KB contains all statements having one of this homonyms in subject or object

EXPERIMENTS

DBPEDIA AND WIKIDATA

True positive	False positive	False Negative	Precision	Recall	F M
33	3	3	0.917	0.917	0

- The 3 false positives are the same than the false negatives
- Each times the right candidates was the second one

EXPERIMENTS

OAEI 2017

Our **goal** is to compare our approach against state of the art approaches that use NLP techniques.

EXPERIMENTS

OAEI 2017

SPIMBENCH SANDBOX: alterations of an original one through value-based, structure-based, and semantics-aware transformations

EXPERIMENTS

OAEI 2017

Participants	Precision	Recall	F-Measure
AML	0.849	1.000	0.918
I-Match	0.854	0.997	0.920
Legato	0.980	0.730	0.840
LogMap	0.938	0.763	0.841
Our approach	0.854	0.996	0.920

EXPERIMENTS

OAEI 2017

- Wrong candidate selection with very similar instances
- Arithmetic mean
- Use both KBs

CONCLUSION

- Fully automatized instance matching approach

https://github.com/PHParis/im_prototype

CONCLUSION

- Fully automatized instance matching approach
- Based on semantics and usage of properties

https://github.com/PHParis/im_prototype

CONCLUSION

- Fully automatized instance matching approach
- Based on semantics and usage of properties
- Recall is good

https://github.com/PHParis/im_prototype

CONCLUSION

- Fully automatized instance matching approach
- Based on semantics and usage of properties
- Recall is good
- Lack of false positive detection/correction

https://github.com/PHParis/im_prototype

CONCLUSION

- Fully automatized instance matching approach
- Based on semantics and usage of properties
- Recall is good
- Lack of false positive detection/correction
- Explore other ways to aggregate the different scores

https://github.com/PHParis/im_prototype

CONCLUSION

- Fully automatized instance matching approach
- Based on semantics and usage of properties
- Recall is good
- Lack of false positive detection/correction
- Explore other ways to aggregate the different scores
- Refine linkset we produced to have fewer false positives results

https://github.com/PHParis/im_prototype

EXPERIMENTS

DBPEDIA AND WIKIDATA

- Source KB contains 277 instances, 3468 triples and 36 candidate instances
- Target KB contains 1170 instances, 7667 triples and 552 candidate instances
- All owl:sameAs links removed and saved as gold standard

EXPERIMENTS

OAEI 2017

- Source KB contains 1432 instances, 10883 triples and 349 candidate instances
- Target KB contains 1453 instances, 10868 triples and 443 candidate instances
- Gold standard provided

WEIGHT OF A ROLE

$$W_{\mathcal{KB}}(R, C) = \frac{NS_{\mathcal{KB}}(C,R)}{NS_{\mathcal{KB}}(C)} \text{ and } W_{\mathcal{KB}}(R, C) \in [0, 1]$$

DISCRIMINATING POWER

$$D_{\mathcal{KB}}(C, R, o) = \frac{NS_{\mathcal{KB}}(C,R,o)}{NS_{\mathcal{KB}}(C,R)} \text{ and } D_{\mathcal{KB}}(C, R, o) \in [0, 1]$$

WEIGHT OF A CLUE

$$\frac{|R_{x_1} \cap R_{x_2}|}{|R_{x_1}| + |R_{x_2}| - |R_{x_1} \cap R_{x_2}|}, \text{ where } clue \in [0, 1]$$