A Network Science approach to Instance Matching

Gabriele Pisciotta[1] and Giulio Rossetti[2]

[1] University of Pisa, Italy g.pisciotta1@studenti.unipi.it
[2] KDD Lab, ISTI-CNR, Italy giulio.rossetti@isti.cnr.it

Nowadays, the use of Knowledge Graphs (KG) as background knowledge is widespread in Machine Learning. Companies and users continuously create and share Linked Open Data: as a result it is common that a same real world entity can be described differently by different KGs. Identifying identity relations is a challenging task, usually called as Instance Matching (IM), that enables inter-linking of different KGs with the aim of supporting the expansion of the knowledge available to ML and Intelligent Agents systems. The IM can be seen as a special case of the Link Prediction (LP) task, in which the only edge type to be predicted between individual of different KGs is the `owl:sameAs`, as shown in Figure 1.

Quoting [2], the IM problem requires to find classifier that effectively discriminate between matching instances and non-matching instances.

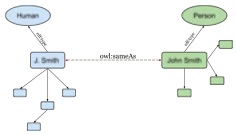

Figure 1: Example of matched instances

	Our Method			CODI		
	F1	P	R	F1	P	R
Value	0.90	0.99	0.85	0.96	0.99	0.93
Structural	0.63	0.72	0.60	0.88	0.95	0.83
Logical	0.92	0.93	0.92	0.97	0.96	0.97
Mixed	0.61	0.75	0.53	0.65	0.86	0.54

Table 1: Results for IIMB 2010 [1]

Considering the network-based nature of the data, we decided to discard any linguistic feature attached to the entities (usually exploited to address the IM task) and to focus on KGs topology alone: for each possible edge between the entities coming from KG pairs, discarding those belonging to disjoint classes, we compute: (i) the Jaccard distance between the two neighbours set, (ii) the Resource Allocation Index , (iii) the Adamic Adar coefficient , (iv) the Preferential Attachment.

We tested our approach on the IIMB 2010 dataset of the Ontology Evaluation Alignment Campaign [1] where the task requires to match entities of an original dataset to the ones of its 80 perturbations embedding various kinds of data transformation (including value, structural, logical, and a mixed ones). We applied AdaBoost to a 80-20 training-test split of the edges for each perturbation sub-task: Table 1 reports the average results for each type of transformation in terms of Precision/Recall and F1. Our results, even with a limited feature set, are in line with SOTA that leverage linguistic features analysis: moreover, our pure topological approach appears robust to value and logical transformation, downgrading its performances only when structural and mixed transformations are applied.

Concluding, our preliminary work underlines how IM can be successfully tackled without involving linguistic features related to the textual description of the entities, while focusing only on Network Science based features. As future works, we plan to extend the experiments to other datasets enriching the feature set and model definition strategies.

References

[1] Jérôme Euzenat et al. “Results of the Ontology Alignment Evaluation Initiative 2010”. In: *Proc. 5th ISWC workshop on ontology matching (OM)*. 2010, pp. 85–117.

[2] Shu Rong et al. “A Machine Learning Approach for Instance Matching Based on Similarity Metrics”. In: *The Semantic Web – ISWC 2012*. 2012, pp. 460–475.