Triple-based Similarity Propagation for Linked Data Matching

Eun-kyung Kim, Sangha Nam, Jongseong Woo, Sejin Nam, and Key-Sun Choi

School of Computing, KAIST, Republic of Korea
`{kekeeo,namsangha,woo88,namsejin,kschoi}@world.kaist.ac.kr`

Abstract. In this paper, we propose an approach for mapping properties in two RDF datasets between different languages, using a triple-based similarity propagation that can be adapted to find potential property matches. This approach does not need any language dependent information during the process, and thus can be applied to arbitrary languages without requiring translation.

1 Introduction

Linked Data aims to extend the Web by publishing various open datasets as RDF and establishing connections between them. DBpedia exploits the huge amount of information contained in Wikipedia and creates a comprehensive dataset by integrating information from many different Wikipedia editions according to an ontology maintained by the community. Due to the interdisciplinary nature and the enormous breadth of coverage of Wikipedia, DBpedia is regarded as one of the central interlinking-hubs of Linked Data [1]. In this paper, we propose an approach for mapping *properties* across DBpedia RDF datasets written in the two languages using a triple-based similarity propagation that can be adapted to find potential property matches without any translation task.

2 Proposed Approach

The proposed approach has two steps: 1) findings the equivalent *subject* and *object* values across datasets at the entity-level, which is represented in the form triples, that are connected by `owl:sameAs` links, and then considering the associated *properties* to have the potential to be equivalent. 2) Then, using a small number of identified matches as seeds to exploit the conceptual-level alignments to identify and estimate semantic relatedness of *properties.* Often, the conceptualizations of triples (from instance triples) are efficient in terms of coverage of alignment, but their result may be dependent on recognizing entities and their type. The types of an entity may not always be present in the dataset. The 'similarity flooding approach' [2] propagates the similarities between concepts to refine the matching results. For example, two apparently different entities from two ontologies are similar when their neighboring concepts are similar.

Experiments: The goal of this experiment is to align language-local properties (i.e., DBpedia Korean property in this case) with the ontological properties of DBpedia in English. Three human annotators aligned 1,000 DBKP to DBOP, if the meaning of two properties was similar. We used the majority vote to determine the correct mapping results.

Table 1. *cf* is the confidence score of the derived property pairs. **I**, **P1**, and **P0** represent the three kinds of propagation scale strategies. **I** denotes cases in which the alignment process is done without using propagation technique. **P1** denotes results obtained from the similarity propagation with the seed with a *cf*=1, whereas cases with a **P0** executes the propagation step with a larger seed with a *cf* >=0. #(**M**) signifies the number of newly discovered matches, and **P**, **R**, and **F** means precisions, recalls and F1-scores, respectively.

	(**I**) w/o prop				(**P1**) prop:seed.θ=1				(**P0**) prop:seed.θ >=0			
cf.	#(**M**)	P	R	F	#(**M**)	P	R	F	#(**M**)	P	R	F
1	13	100	0.96	1.91	25	100	0.96	1.91	23	95.65	1.3	2.57
0.9	42	95.24	2.97	5.76	50	94.44	2.52	4.91	56	94.64	3.14	6.07
0.7	98	96.94	7.05	13.14	118	93.26	6.16	11.55	121	97.52	6.99	13.04
0.5	151	96.69	10.83	19.48	226	90.70	11.57	20.53	199	97.99	11.55	20.66
0.4	188	95.74	13.35	23.44	282	91.12	14.47	24.97	246	95.93	13.97	24.39
0.3	222	95.5	15.73	27.01	386	88.66	19.14	31.48	306	94.12	17.05	28.87
0.2	269	94.42	18.84	31.42	538	82.80	22.85	35.81	381	89.5	20.19	32.95
0.1	322	92.55	22.11	35.69	863	76.01	27.97	40.89	505	84.95	25.4	39.11
0	668	75.15	37.24	49.80	3,166	59.35	47.11	52.52	896	74.33	39.43	51.53

Analysis: The preliminary experiment between the English and the Korean DBpedia has shown that the propagated connectives improve the recall and F1-score measures required to find mapping pairs of properties by taking into account instance types in order to discover new mapping candidates. We see this as the initial step towards enhancing multilingualism in Linked Open Data.

Acknowledgement This work was supported by Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIP) (No. R0101-15-0054, WiseKB: Big data based self-evolving knowledge base and reasoning platform)

References

1. Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P.N., Hellmann, S., Morsey, M., van Kleef, P., Auer, S., Bizer, C.: DBpedia - a large-scale, multilingual knowledge base extracted from wikipedia. Semantic Web Journal (2014)
2. Melnik, S., Garcia-Molina, H., Rahm, E.: Similarity flooding: A versatile graph matching algorithm and its application to schema matching. In: Agrawal, R., Dittrich, K.R. (eds.) ICDE. pp. 117–128. IEEE Computer Society (2002), http://dblp.uni-trier.de/db/conf/icde/icde2002.html#MelnikGR02