Types of Property Pairs and Alignment on Linked Datasets – A Preliminary Analysis

Kalpa Gunaratna, Krishnaprasad Thirunarayan, and Amit Sheth

Kno.e.sis Center, Wright State University, Dayton OH, USA
{kalpa,tkprasad,amit}@knoesis.org
http://knoesis.org

Abstract. Dataset publication on the Web has been greatly influenced by the Linked Open Data (LOD) project. Many interlinked datasets have become freely available on the Web creating a structured and distributed knowledge representation. Analysis and aligning of concepts and instances in these interconnected datasets have received a lot of attention in the recent past compared to properties. We identify three different categories of property pairs found in the alignment process and study their relative distribution among well known LOD datasets. We also provide comparative analysis of state-of-the-art techniques with regard to different categories, highlighting their capabilities. This could lead to more realistic and useful alignment of properties in LOD and similar datasets.

Keywords: Linked Data, Property Alignment, Property Pair Analysis

1 Introduction

LOD [2] has popularized the way individual datasets can be published on the Web by making inter-connections. This has resulted in the creation of a huge structured knowledge graph on the Web. Since dataset publishers are autonomous and design their datasets to meet their respective purposes for originally developing datasets, data interoperability and data integration tasks on these datasets are challenging. Property alignment is one such research problem where innovative solutions are required to handle complex data representations in these interconnected datasets that go well beyond simple string manipulations.

We introduced a novel way of computing property alignment (similarity) between interconnected datasets by exploring the available links between the datasets and using statistical measures [4]. Our solution can successfully handle complex data representations found at the property level in the matching process. We start with a breakdown of types of property pairs found on the LOD and discuss the performance of matching algorithms on the non-trivial task of property alignment between datasets. The analysis is based on manually identified and categorized property pairs of a sample of well known linked datasets in the LOD cloud. Moreover, the analysis presents how many of the manually identified property pairs in each category are identified by the different matching techniques (recall for each property type) highlighting their applicability.

2 Analysis

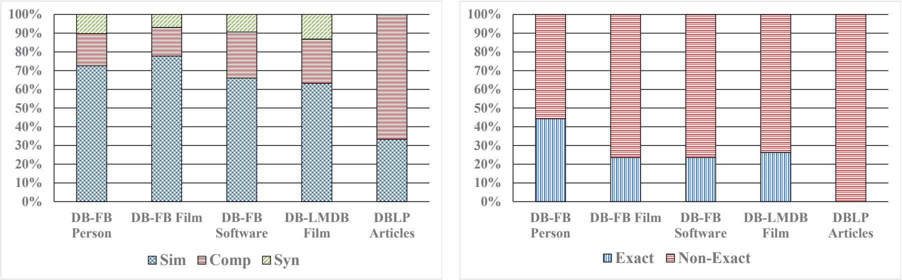
We analyze different types of property pairs found along with the experiments performed in [4]. Such analysis can provide a deeper understanding of types of property pairs that exist in linked datasets and how matching of such property pairs can be improved between two linked datasets using property extensions.

We can categorize the types of property pairs between linked datasets in two orthogonal ways: (1) on the basis of their semantics, and (2) on the basis of the techniques and tools required to determine the inter-relationships or alignment among property pairs. On the basis of semantics, the related property pairs can be classified as (1) equivalent properties or (2) those possessing a property-sub property relationship. On the basis of the techniques used to align properties, we can classify property pairs as follows:

1. *Simple property pairs*: These have high syntactic similarity in the property names and may have a common prefix, common suffix, adjectives, or different ordering of words, e.g., *birthPlace* vs *placeOfBirth*. Here the words "place" and "birth" are in a different order for the two properties.
2. *Opaque property pairs*: These have the same meaning but use different words. This can be further categorized into two parts.
   (a) *Synonymous property pairs*: Similarity of the two properties can be decided by analyzing the meaning of the property names and is intentional. This can be achieved by using an external dictionary or a lexical database like WordNet. If property name is a word phrase, similarity can be checked by removing common words from the property names, e.g., *occupation* vs *profession*, *city of birth* vs *place of birth*. In the second property pair, the common suffix can be eliminated from the comparison.
   (b) *Complex property pairs*: Similarity cannot be determined by considering property names alone, but requires additional information such as extension analysis, and domain and range. These are ambiguous or have multiple meanings but have a specific meaning in a dataset, e.g., *battle* vs *participated in conflict*, *resting place* vs *place of burial*. The two terms "conflict" and "resting place" have multiple meanings and are used in many contexts. Hence, the similarity is harder to identify.

In this analysis, we highlight the advantages of using property extensions compared to string based and external dictionary based methods that focus on analyzing property names in the matching process. We consider only object-type properties for this analysis in DBpedia, Freebase, LinkedMDB, and DBLP datasets[1], taking 5000 instances in each sample set [1]. We did not consider property chains or composite property alignment in this preliminary analysis, which belong to the complex property pair type. Composite property alignment is the process of aligning a property in one dataset with several properties (or property chains) in another dataset. There exist other efforts (within datasets), different from ours, that analyze sets of properties in RDF [6], combination of properties and classes in LOD [3], and time dynamics of LOD [5].

[1] person, film and software domains between DBpedia and Freebase, films between DBpedia and LinkedMDB, and articles in DBLP (L3S and RKB Explorer).



(a) Property pair type breakdown

(b) Exact matching of property pairs

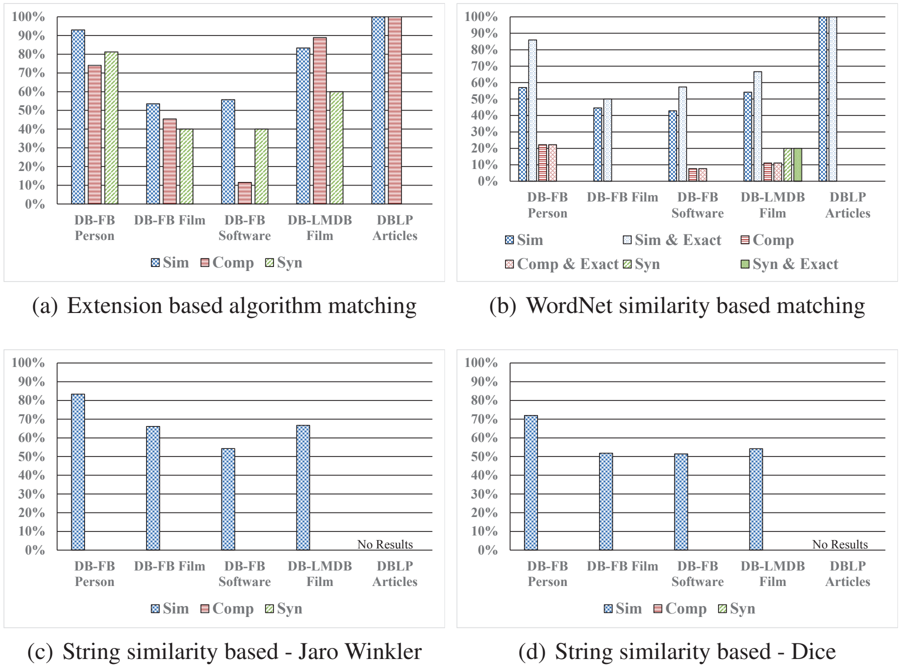
Fig. 1. Property pairs breakdown. Syn for Synonymous, Comp for Complex and Sim for Simple property pairs.

The correct matches in this analysis were manually identified and categorized by the authors and verified by an external reviewer. Figure 1 shows the breakdown of properties into the three types that we are interested in. According to Figure 1(a), the majority of the property pairs belong to simple property pairs followed by complex and synonymous property pairs. Moreover, some property pairs can be matched using exact property name matching as shown in Figure 1(b), but they account for less. Based on the facts presented in Figure 1, on average, the majority of the matching property pairs are simple, but cannot be matched using exact matching of property names.

There are different approaches for aligning property pairs between datasets including [4], which is based on property extension matching. In the extension based approach, alignment of two properties is decided by aggregating the number of matched subject-object pairs in the property extension over the number of co-appearances of the property pair in two linked datasets. We utilized Entity Co-Reference (ECR) links that exist between linked datasets in matching extensions. That is, two instances (in the property extension) are considered the same if they are connected by an *ECR* link. There can be incorrect matches for each property as extensions of properties overlap. For example, ”birthPlace” property may match to ”deathPlace” with some overlap in the extension, but when the whole result set is aggregated and analyzed, these coinsidental matches can be eliminated. For the WordNet based approach, we calculated the normalized WordNet similarity using eight similarity measures[2] found in the literature over terms appearing in the property names after removing stop words. For string similarity measurements, we added stemming in the preprocessing step before computing the similarity over property names. More details including threshold values and formulas used for matching are in [4][1].

Considering these matchers, Figure 2 shows the percentages of the correctly identified property pairs for the three types of property pairs. It also shows the superiority of the extension based approach over string based and dictionary (WordNet) based approaches. It is clear from Figures 2(a), 2(b), 2(c), and 2(d) that the extension based

[2] namely, LCH, RES, HSO, JCN, LESK, PATH, WUP and LIN



(a) Extension based algorithm matching

(b) WordNet similarity based matching

(c) String similarity based - Jaro Winkler

(d) String similarity based - Dice

Fig. 2. Matching % (recall) for each type of property pair using different approaches. Syn for Synonymous, Comp for Complex and Sim for Simple property pairs.

approach performed better and achieved the highest results in matching all three types of property pairs. We added exact matching of property names capability to WordNet based algorithm and improved its performance as shown in Figure 2(b). This is because some word phrases cannot be matched (searched) using WordNet but they have the same or common word phrases in their names. It is also interesting to note that the WordNet based approach failed to identify any of the synonymous property pairs in most of the experiments as shown in Figure 2(b). This kind of behavior is expected for string similarity or syntax based approaches, but not for a lexical database based approach like WordNet, which is specialized in synonym word categorization. Figures 2(c) and 2(d) present matching performances when the similarity of property names are considered using string matching algorithms. It is shown that string similarity based matching missed all synonymous and complex property pairs leaving them unsuitable for matching property pairs in general. Based on the facts (recall values) represented in Figure 2, extension based property alignment has the capability to identify many property pairs including complex and hidden property pairs compared to others. Furthermore, Table 1 outlines both precision and recall for each matcher for all property pair types, which also sheds lights on false positives (see [4] for more details). Note that it is not possible to provide a precision value breakdown for each property pair type, since we are not identifying each type in the alignment process but all.

	Measure Type	DBpedia-Freebase (Person)	DBpedia-Freebase (Film)	DBpedia-Freebase (Software)	DBpedia-LinkedMDB (Film)	DBLP_RKB-DBLP_L3S (Article)	Average
Extension Based Algorithm	Precision	**0.8758**	**0.9737**	0.6478	0.7560	**1.0000**	**0.8427**
	Recall	**0.8089***	**0.5138**	**0.4339**	**0.8157**	**1.0000**	**0.7145**
	F measure	**0.8410***	**0.6727**	**0.5197**	**0.7848**	**1.0000**	**0.7656**
Dice Similarity	Precision	0.8064	0.9666	0.7659	**1.0000**	0.0000	0.7078
	Recall	0.4777*	0.4027	0.3396	0.3421	0.0000	0.3124
	F measure	0.6000*	0.5686	0.4705	0.5098	0.0000	0.4298
Jaro Similarity	Precision	0.6774	0.8809	**0.7755**	0.9411	0.0000	0.6550
	Recall	0.5350*	**0.5138**	0.3584	0.4210	0.0000	0.3656
	F measure	0.5978*	0.6491	0.4903	0.5818	0.0000	0.4638
WordNet Similarity	Precision	0.5200	0.8620	0.7619	0.8823	**1.0000**	0.8052
	Recall	0.4140*	0.3472	0.3018	0.3947	0.3333	0.3582
	F measure	0.4609*	0.4950	0.4324	0.5454	0.5000	0.4867

Table 1. Alignment of object-type properties. Boldface and * mark highest and estimated values.

3 Conclusion

We provided a breakdown of types of property pairs that can be found on linked datasets in the alignment process. Even though the majority of the property pairs are simple, many cannot be identified using string manipulation techniques. In our sample datasets, 63%, 29%, and 8% of all property pairs are simple, complex, and synonymous, respectively. We have shown that in every category, extension based property pair alignment showed better results. For example, the extension based approach showed an average improvement in the range of 5% - 32% compared to simple syntactic and WordNet based approaches. Hence, we conclude that the extension (or instance) based approach can discover many property pairs that are semantically the same, which cannot be uncovered by purely syntactic means.

Acknowledgement - This work was supported by the NSF award 1143717 “III: EAGER Expressive Scalable Querying over Linked Open Data”. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

1. More at, `http://wiki.knoesis.org/index.php/Property_Alignment`
2. Bizer, C., Heath, T., Berners-Lee, T.: Linked data-the story so far. International Journal on Semantic Web and Information Systems (IJSWIS) 5(3), 1–22 (2009)
3. Gottron, T., Knauf, M., Scheglmann, S., Scherp, A.: A systematic investigation of explicit and implicit schema information on the linked open data cloud. In: The Semantic Web: Semantics and Big Data, pp. 228–242. Springer (2013)
4. Gunaratna, K., Thirunarayan, K., Jain, P., Sheth, A., Wijeratne, S.: A statistical and schema independent approach to identify equivalent properties on linked data. In: 9th International Conference on Semantic Systems. ACM (2013)
5. Käfer, T., Abdelrahman, A., Umbrich, J., OByrne, P., Hogan, A.: Observing linked data dynamics. In: The Semantic Web: Semantics and Big Data, pp. 213–227. Springer (2013)
6. Neumann, T., Moerkotte, G.: Characteristic sets: Accurate cardinality estimation for rdf queries with multiple joins. In: Data Engineering (ICDE), 2011 IEEE 27th International Conference on. pp. 984–994. IEEE (2011)