Discovering Alignments in Ontologies of Linked Data*

Rahul Parundekar, Craig A. Knoblock, and José Luis Ambite
University of Southern California
Information Sciences Institute
4676 Admiralty Way
Marina del Rey, CA 90292

Abstract

Recently, large amounts of data are being published using Semantic Web standards. Simultaneously, there has been a steady rise in links between objects from multiple sources. However, the ontologies behind these sources have remained largely disconnected, thereby challenging the interoperability goal of the Semantic Web. We address this problem by automatically finding alignments between concepts from multiple linked data sources. Instead of only considering the existing concepts in each ontology, we hypothesize new composite concepts, defined using conjunctions and disjunctions of (RDF) types and value restrictions, and generate alignments between them. In addition, our techniques provide a novel method for curating the linked data web by pointing to likely incorrect or missing assertions. Our approach provides a deeper understanding of the relationships between linked data sources and increases the interoperability among previously disconnected ontologies.

1 Introduction

The last few years have witnessed a paradigm shift from publishing isolated data to publishing data that is *linked* to related data from other sources using the structured model of the Semantic Web. By doing so, the publishers of the linked data are able to supplement their own knowledge base, by integrating data from different sources, and realize significant benefits across various domains. Most of the effort has been on identifying which objects from different sources are actually the same. For example, that object *geonames.org/5368361* is the same as *dbpedia:Los_Angeles*. Despite the increasing availability of linked data, the absence of links at the concept level has resulted in heterogenous schemas, challenging the interoperability goal of the Semantic Web. For example, of the 190 sources in the latest census of linked data[1] only 15 have mappings between their ontologies.

The problem of schema linking (aka schema matching in databases and ontology alignment in the Semantic Web) has received much attention [Bellahsene *et al.*, 2011; Euzenat and Shvaiko, 2007; Bernstein *et al.*, 2011; Gal, 2011]. In this paper we present a novel extensional approach to generate alignments between ontologies of linked data sources. Similar to previous work on instance-based matching [Duckham and Worboys, 2005; Doan *et al.*, 2004; Isaac *et al.*, 2007], we rely on linked instances to determine the alignments. Two concepts are equivalent if all (or most of) their respective instances are linked (by *owl:sameAs* or similar links). However, our search is not limited to the existing concepts in the ontology. We hypothesize new concepts by combining existing elements in the ontologies and seek alignments between these more general concepts. This ability to generalize allows us to find many more meaningful alignments in ontologies in which one-to-one concept equivalences might not exist. For example, the alignment of an impoverished ontology like *GeoNames*, which has only one class - *geonames:Feature*, with the well-developed *DBpedia* ontology is not particularly informative. To successfully link such ontologies, we first generate more expressive concepts, based on properties and values of the instances in the sources. For example, in *GeoNames* the values of the *featureCode* and *featureClass* properties can be used to find alignments with existing concepts in *DBpedia*, such as the alignment of the concept *geonames:featureClass=P* to *dbpedia:PopulatedPlace*.

Our approach finds alignments between concepts defined by conjunction and disjunctions of (RDF) type and value restrictions (cf. [Horrocks *et al.*, 2006]), which we call *restriction classes* henceforth. An *atomic restriction class*, $\{p = v\}$, is the set of objects having object (or data) property p (including rdf:type) with object (or literal) value v. These alignments are based on the linked instances between these composite concepts. This is an important feature of our approach; we model the *actual* contents and relationships between sources, as opposed to what ontologies disassociated from the data may lead us to believe based on class names or structure.

2 Sources with Heterogenous Ontologies

Linked data sources often conform to different, but related, ontologies that can be meaningfully linked [Cruz *et al.*, 2011; Jain *et al.*, 2011; Parundekar *et al.*, 2010; 2012]. Our algorithms are generic and can be used to align any two linked

---

*The paper on which this extended abstract is based was the recipient of the best paper award in the research track at the 11th International Semantic Web Conference, 2012 [Parundekar *et al.*, 2012].

[1] http://www4.wiwiss.fu-berlin.de/lodcloud/state/

sources. However, we will use two sources with geospatial data for better illustration of our approach. *GeoNames* (geonames.org), contains about 7.8 million geographical objects. It is described by a rudimentary ontology since its semantic web version was generated automatically by direct translation of a simple relational database model. All its instances belong to a single class, *Feature*, with the type of the geographical data (e.g. mountains, lakes, cities, etc.) encoded in the *featureClass* and *featureCode* properties. *DBpedia* (dbpedia.org) is a knowledge base that covers multiple domains and includes approximately 526,000 geographical objects. It is described using a rich ontology with extensive concept hierarchies and numerous relations. At the time of our experiments, these two sources have over 86,000 pairs of instances linked using *owl:sameAs* assertions.

3 Finding Alignments Across Ontologies

We find three types of alignments between the ontologies of linked data sources. First, we extract equivalent and subset alignments between *atomic restriction classes*. These are the simplest alignments that we define. Though simple, they often yield interesting alignments. Moreover, we use them as seed hypotheses to find alignments that are more descriptive. Second, we find alignments between *conjunctive restriction classes* in the two sources. Finally, we find *concept coverings*, which are alignments where a concept from one source maps to a union of smaller concepts from the other source.

Before searching for alignments, we pre-process the sources to reduce the search space and avoid computation not leading to meaningful alignments. First, we only consider instances that are actually linked, thus removing unrelated instances and their properties. Second, we eliminate inverse (or quasi-inverse) functional properties, since a *restriction class* on such a property would only contain a single instance (or very few) and would not be a useful concept (e.g., the latitude and longitude properties generally point to one place).

3.1 Aligning Atomic Restriction Classes

Atomic restriction classes can be generated by combining properties and values in the sources and tested for alignments using the simple algorithm in Fig. 1. Fig. 2 illustrates the set comparison operations of our algorithm. We consider the two concepts equivalent if they significantly overlap each other. We use two metrics *P* and *R* to measure the degree of overlap between *restriction classes*. In a perfect equivalence alignment, the values for P and R would be both 1. However, to allow for missing links or errors in the sources, we use $P \geq \theta$ and $R \geq \theta$ ($\theta = 0.9$ in our experiments). For example, consider the alignment between *restriction classes* {*geonames:countryCode*=ES} and {*dbpedia:country* = *dbpedia:Spain*}. Based on the concept extensions, our algorithm finds $|Img(r_1)|$ = 3198, $|r_2|$ = 4143, $|Img(r_1) \cap r_2|$ = 3917, R = 0.9997 and P = 0.9454. Thus, the algorithm considers this alignment as equivalent in an extensional sense. This algorithm finds numerous equivalent and subset alignments between *atomic restriction classes*. For example, we find that each of {*geonames:featureCode* = *S.SCH*} and {*geonames:featureCode* = *S.UNIV*} (i.e. Schools and Universities from *GeoNames*) are subsets of {*dbpedia:EducationalInstitution*}.

function ATOMICALIGNMENTS($Source_1$,$Source_2$)
  for all properties p_1 in $Source_1$, **all** distinct values $v_1 \in p_1$, **all** p_2 in $Source_2$, and **all** distinct $v_2 \in p_2$ **do**
    $r_1 \leftarrow \{p_1 = v_1\}$ // instances of $Source_1$ with $p_1 = v_1$
    $r_2 \leftarrow \{p_2 = v_2\}$
    $Img(r_1) \leftarrow$ instances of $Source_2$ linked to those in r_1
    $P \leftarrow \frac{|Img(r_1) \cap r_2|}{|r_2|}$, $R \leftarrow \frac{|Img(r_1) \cap r_b|}{|r_1|}$
    $alignment(r_1, r_2) \leftarrow$ [
    if $P \geq \theta$ *and* $R \geq \theta$ **then** $r_1 \equiv r_2$
    else if $P \geq \theta$ **then** $r_1 \subset r_2$
    else if $R \geq \theta$ **then** $r_2 \subset r_1$
    end if]
  end for
end function

Figure 1: Aligning *atomic restriction classes*

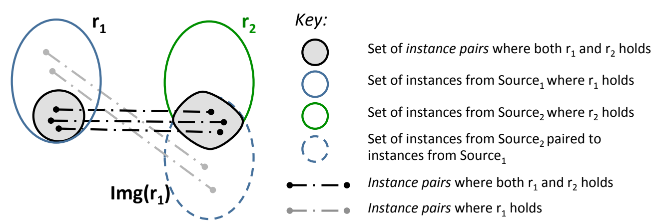

Figure 2: Comparing linked instances from two ontologies

3.2 Aligning Conjunctive Restriction Classes

The second type of alignments we detect are those between *conjunctive restriction classes*. For example, the *conjunctive restriction class* 'Schools in the US', {*geonames:countryCode=US* $\cap$ *geonames:featureCode=S.SCH*}, is the intersection of the *atomic restriction classes* representing all schools in *GeoNames* and all things in the US.

We seed the search space with the alignments generated by ATOMICALIGNMENTS. Taking one hypothesis at a time, we can generate new hypothesis by conjoining *atomic restriction classes*. For example, we can extend the alignment [{*geonames:featureCode=S.SCH*}, {*rdf:type=EducationalInstitution*}] by conjoining {*geonames:featureCode=S.SCH*} with {*geonames:countryCode=US*}, and investigate the relationship between schools in the US and educational institutions.

The algorithm to find *conjunctive restriction classes* appears in Fig. 3 (cf. [Parundekar *et al.*, 2010]). To reduce the combinatorial search space, our algorithm prunes hypotheses that 1) do not have enough instances supporting either of the *restriction classes*; 2) where the extension of the refined *restriction class* (r') is the same as its parent (i.e. the constraint did not specialize the concept); 3) have the form [r'_1, r_2], where r'_1 is a subclass of r_1 and $r_1 \subset r_2$, since no immediate specialization is provided; and 4) would have been explored more than once (these are avoided by using a lexicographic ordering). Finally, the algorithm removes any implied alignments that are generated because of the hierarchical nature of the algorithm. Specifically, it removes 1) relations that implied by the transitivity of the subset relations; 2) cyclic equivalences which may be generated due to the reduced threshold θ.

function CONJUNCTIVEALIGNMENTS($Source_1$,$Source_2$)
  for all $[r_1, r_2] \in$ ATOMICALIGNMENTS($Source_1$,$Source_2$)
do EXPLOREHYPOTHESES(r_1, r_2,$Source_1$,$Source_2$)
  end for
  REMOVEIMPLIEDALIGMENTS
end function
function EXPLOREHYPOTHESIS(r_1,r_2,$Source_a$,$Source_b$)
  for all p_a in $Source_a$ occurring lexicographically after all the properties in r_1 and distinct v_a associated with p_a **do**
    $r_1' \leftarrow r_1 \cap \{p_a = v_a\}$
    $alignment \leftarrow$ FINDALIGNMENT(r_1',r_2)
    if not SHOULDPRUNE(r_1',r_2,$alignment$) **then**
      $alignment(r_1', r_2) \leftarrow alignment$
      EXPLOREHYPOTHESES(r_1', r_2)
    end if
  end for
  for all p_b in $Source_b$ occuring lexicographically after all the properties in r_2 and distinct v_b associated with p_b **do**
    $r_2' \leftarrow r_2 \cap \{p_b = v_b\}$
    $alignment \leftarrow$ FINDALIGNMENT(r_1,r_2')
    if not SHOULDPRUNE(r_1,r_2',$alignment$) **then**
      $alignment(r_1, r_2') \leftarrow alignment$
      EXPLOREHYPOTHESES(r_1, r_2')
    end if
  end for
end function

Figure 3: Aligning *conjunctive restriction classes*

3.3 Finding Concept Coverings

The CONJUNCTIVEALIGNMENTS algorithm may produce a very large number of subset relations. Analyzing the results of [Parundekar *et al.*, 2010], we noticed that these subset alignments follow common patterns. For example, we found that both Schools and Universities from *GeoNames* were subsets of Educational Institutions from *DBpedia*. However, the *union* of Schools, Colleges, and Universities from *GeoNames* was *equivalent* to *dbpedia:EducationalInstitution*, which is a more informative finding.

The algorithm for generating *concept coverings* appears in Fig. 4. We start with the subclass alignments found by ATOMICALIGNMENTS. Then we identify concepts from one ontology that are defined on the same property and are subsets of another concept in the other ontology. We test whether the union of the smaller concepts is equivalent to the larger concept based on the extensions of the concepts as before. Although we could explore more complex hypotheses, this approach is tractable and generates intuitive definitions.

Since all smaller classes are subsets of the larger *restriction class*, $P_U \geq \theta$ holds by construction. Thus, we just need to check that $R_U \geq \theta$ to determine whether the union *restriction class* is equivalent to the single concept. The smaller *restriction classes* that were omitted in ATOMICALIGNMENTS because of insufficient support size of their intersections (e.g.,

function CONCEPTCOVERINGS($Source_1$,$Source_2$)
  for all alignments $[U_L, r_2] \in$ ATOMICALIGNMENTS($Source_1$,$Source_2$), with larger concept $U_L = \{p_L = v_L\}$ from $Source_1$ and multiple classes $r_2 = \{p_S = v_i\}$ from $Source_2$ that can be partitioned on property p_S **do**
    for all smaller concepts $\{p_S = v_i\}$ **do**
      $U_S \leftarrow \{p_S = \{v_1, v_2, ...\}\}$ // union *restriction class*
      $U_A \leftarrow Img(U_L) \cap U_S$, $P_U \leftarrow \frac{|U_A|}{|U_S|}$, $R_U \leftarrow \frac{|U_A|}{|U_L|}$
      if $R_U \geq \theta$ **then** $alignment(r_1, r_2) \leftarrow U_L \equiv U_S$
      end if
    end for
  end for
end function

Figure 4: Finding Concept Coverings

{*geonames:featureCode = S.SCHC*}) are included in constructing U_S for completeness.

Figure 5 illustrates the approach. ATOMICALIGNMENTS detects that {*geonames:featureCode = S.SCH*}, {*geonames:featureCode = S.SCHC*}, *and* {*geonames:featureCode = S.UNIV*} are subsets of {*rdf:type = dbpedia:EducationalInstitution*}. As can be seen in the Venn diagram in Figure 5, U_L is *Img*({*rdf:type = dbpedia:EducationalInstitution*}), U_S is {*geonames:featureCode = S.SCH*} ∪ {*geonames:featureCode = S.SCHC*} ∪ {*geonames:featureCode = S.UNIV*}, and U_A is the intersection of the two. Upon calculation we find that R_U for the alignment of *dbpedia:EducationalInstitution* to {*geonames:featureCode*= {S.SCH, S.SCHC, S.UNIV}} is 0.98 (greater than θ). We can thus confirm the hypothesis and consider U_L and U_S as equivalent.

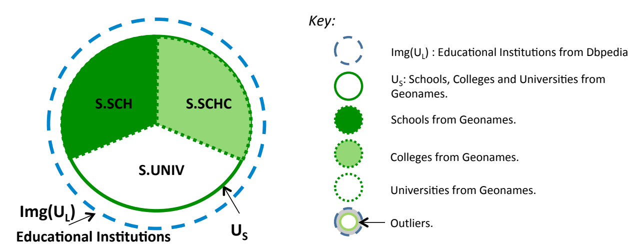

Figure 5: Concept covering: Educational Institutions

4 Results

The results of the three alignment algorithms over *GeoNames* and *DBpedia* appear in Table 2. In all, we were able to detect about 580 (263 + 45 + 221 + 51) equivalent alignments including both atomic and complex *restriction classes*, along with 15,376 (4,946 + 5,494 + 4,400 +536) subset relations.

Table 1 shows some of the representative alignments found by the three algorithms. We are able to detect alignments of *atomic restriction classes* with existing RDF concepts (i.e., defined with *rdf:type*) and with value restrictions. For example, alignment #1 shows that the feature class ‘H’ in *GeoNames* maps to a ‘Body of Water’ in *DBpedia*. Alignments #2 and #3 show equivalence and subset relations between value restrictions. Alignment #4 shows a conjunctive alignment for ‘Populated Places in the US’. Finally, alignment #5

#	*GeoNames* concept	Rel.	*DBpedia* concept	P	R	$\|I(r_1) \cap r_2\|$
1	geonames:featureClass=geonames:H	$=$	rdf:type=dbpedia:BodyOfWater	0.91	0.99	1939
2	geonames:countryCode=ES	$=$	dbpedia:country=dbpedia:Spain	0.95	0.99	3917
3	geonames:featureCode=geonames:T.MT	$\subset$	rdf:type=dbpedia:Mountain	0.97	0.78	1721
4	geonames:featureClass=geonames:P & geonames:countryCode=US	$=$	rdf:type=dbpedia:PopulatedPlace & dbpedia:country=dbpedia:United_States	0.97	0.96	26061
5	*geonames:featureCode* = {S.SCH, S.SCHC, S.UNIV}	=	*rdf:type* = *dbpedia:EducationalInstitution*	-	0.98	396

Table 1: Representative alignments found in two sources

atomic restriction classes	**Alignments**
Equivalent Alignments	263
Subclasses with larger class from *GeoNames*	4,946
Subclasses with larger class from *DBpedia*	5,494
conjunctive restriction classes	
Equivalent Alignments	45
Subclasses with larger class from *GeoNames*	4,400
Subclasses with larger class from *DBpedia*	536
concept coverings	
Coverings with larger class from *GeoNames*	221
Coverings with larger class from *DBpedia*	51

Table 2: Alignments found between *GeoNames* and *DBpedia*

shows the covering of 'Educational Institutions' in *DBpedia* with schools, colleges and universities in *GeoNames*. None of these alignments could be detected by previous algorithms that do not hypothesize concepts beyond the existing classes. Also, note that the alignments generated from *actual* data need not match the intentional similarity of the concepts. For example, Mountains from *GeoNames* are subset of Mountains in *DBpedia*, since *GeoNames* divides the concept by distinguishing Peaks, Hills, etc., from Mountains.

An interesting outcome of our approach is the detection of outliers, which suggest possible erroneous links or value assignments. For example, in alignment #2, R overlap (0.99) is not complete (1) since one outlier instance from *GeoNames* has 'IT' (Italy) as *countryCode*. However, this is likely an error since there is overwhelming support for 'ES' being the *countryCode* of Spain. Alignment #5 shows an interesting case where 8 instances could not be identified as Educational Institutions. They had either a missing *genomes:featureCode* (1) or a value for Library (1), Hospitals (1), Buildings (3), Establishments (1), and Museums (1). The detection of outliers provides a unique opportunity for identifying inconsistencies and automatically curate the web of linked data.

5 Related Work

Even though most previous work on linked data focuses on linking instances across different sources, several authors have considered aligning ontologies of linked data sources. Jain et al. [2010] describe the BLOOMS approach, which uses a central forest of concepts derived from topics in Wikipedia. It is, however, unable to find alignments because of the single *Feature* class in *GeoNames*. BLOOMS+ [Jain *et al.*, 2011] aligns linked data ontologies with an upper-level ontology called Proton. Using contextual information, BLOOMS+ finds an greater number of alignments between *GeoNames* & Proton and *DBpedia* & Proton than its predecessor. Cruz et al. [2011] describe a dynamic ontology mapping approach called *AgreementMaker* that uses similarity measures along with a mediator ontology to find mappings using the labels of the classes. The advantage of our approach is that, by hypothesizing novel concepts (*restriction classes*), it can find a larger set of alignments than previous approaches, even from sources described using a rudimentary ontology, such as *GeoNames*. Völker et al. [2011] describe an extensional approach that uses statistical methods for finding alignments by generating OWL-2 axioms using an intermediate associativity table of instances and concepts and mining associativity rules from it. GLUE [Doan *et al.*, 2004] is a instance-based matching algorithm, which predicts the concept in the other source that instances belong to by using machine learning. GLUE then hypothesizes alignments based on the probability distributions obtained from the classifications. In contrast, our approach is based on the existing links, and hence reflects the nature of the source alignments in practice. *CS*R [Spiliopoulos *et al.*, 2008] aligns a concept from one ontology to a union of concepts from another ontology using the similarity of properties as features in predicting the subsumption relationships. It differs from our approach in that it uses a statistical machine learning approach for detection of subsets rather than the extensional approach. Atencia et al., [2012] provide a formalization of weighted ontology mappings that is applicable to extensional matchers like ours.

6 Conclusion

We described an approach to identifying alignments between atomic, conjunctive and disjunctive *restriction classes* in linked data sources. Our approach discovers alignments where concepts at different levels in the ontologies of two sources can be mapped even when there is no direct equivalence or only rudimentary ontologies exist. Our algorithm is also able to detect outliers that help identify erroneous links or inconsistencies in the linked instances. By using the *GeoNames* and *DBpedia* sources as an example, we showed that the results the algorithm generates can provide a deeper insight into the nature of the alignments of linked data.

In future work, we plan to explore more expressive concept descriptions and provide a curation system that not only signals outliers, but also proposes corrections automatically.

Acknowledgements

This research is based upon work supported in part by the National Science Foundation under award number IIS-1117913. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of NSF or the U.S. Government.

References

[Atencia *et al.*, 2012] Manuel Atencia, Alexander Borgida, Jrme Euzenat, Chiara Ghidini, and Luciano Serafini. A formal semantics for weighted ontology mappings. In *Proceedings of the 11th International Semantic Web Conference (ISWC 2012)*, pages 17–33, Boston, Massachusetts, 2012.

[Bellahsene *et al.*, 2011] Zohra Bellahsene, Angela Bonifati, and Erhard Rahm. *Schema Matching and Mapping*. Springer, 1st edition, 2011.

[Bernstein *et al.*, 2011] P.A. Bernstein, J. Madhavan, and E. Rahm. Generic schema matching, ten years later. *Proceedings of the VLDB Endowment*, 4(11), 2011.

[Cruz *et al.*, 2011] I.F. Cruz, M. Palmonari, F. Caimi, and C. Stroe. Towards on the go matching of linked open data ontologies. In *Workshop on Discovering Meaning On The Go in Large Heterogeneous Data*, page 37, 2011.

[Doan *et al.*, 2004] A.H. Doan, J. Madhavan, P. Domingos, and A. Halevy. Ontology matching: A machine learning approach. *Handbook on ontologies*, pages 385–404, 2004.

[Duckham and Worboys, 2005] M. Duckham and M. Worboys. An algebraic approach to automated geospatial information fusion. *International Journal of Geographical Information Science*, 19(5):537–558, 2005.

[Euzenat and Shvaiko, 2007] Jerome Euzenat and Pavel Shvaiko. *Ontology matching*. Springer-Verlag, 2007.

[Gal, 2011] Avigdor Gal. *Uncertain Schema Matching*. Synthesis Lectures on Data Management. Morgan & Claypool Publishers, 2011.

[Horrocks *et al.*, 2006] Ian Horrocks, Oliver Kutz, and Ulrike Sattler. The even more irresistible SROIQ. In *Proceedings, Tenth International Conference on Principles of Knowledge Representation and Reasoning*, pages 57–67, Lake District of the United Kingdom, June 2006.

[Isaac *et al.*, 2007] A. Isaac, L. Van Der Meij, S. Schlobach, and S. Wang. An empirical study of instance-based ontology matching. *The Semantic Web*, pages 253–266, 2007.

[Jain *et al.*, 2010] P. Jain, P. Hitzler, A. Sheth, K. Verma, and P. Yeh. Ontology alignment for linked open data. *The Semantic Web–ISWC 2010*, pages 402–417, 2010.

[Jain *et al.*, 2011] P. Jain, P. Yeh, K. Verma, R. Vasquez, M. Damova, P. Hitzler, and A. Sheth. Contextual ontology alignment of lod with an upper ontology: A case study with proton. *The Semantic Web: Research and Applications*, pages 80–92, 2011.

[Parundekar *et al.*, 2010] Rahul Parundekar, Craig A. Knoblock, and José Luis Ambite. Linking and building ontologies of linked data. In *Proceedings of the 9th International Semantic Web Conference (ISWC 2010)*, Shanghai, China, 2010.

[Parundekar *et al.*, 2012] Rahul Parundekar, Craig A. Knoblock, and José Luis Ambite. Discovering concept coverings in ontologies of linked data sources. In *Proceedings of the 11th International Semantic Web Conference (ISWC 2012)*, pages 427–443, Boston, Massachusetts, 2012.

[Spiliopoulos *et al.*, 2008] V. Spiliopoulos, A. Valarakos, and G. Vouros. Csr: discovering subsumption relations for the alignment of ontologies. *The Semantic Web: Research and Applications*, pages 418–431, 2008.

[Völker and Niepert, 2011] J. Völker and M. Niepert. Statistical schema induction. *The Semantic Web: Research and Applications (ESWC 2011)*, pages 124–138, 2011.