A Journey From Simple to Complex Alignment on Real-World Ontologies

Lu Zhou

DaSe Lab, Wright State University, Dayton OH 45435, USA
zhou.34@wright.edu

Abstract. Ontology alignment has been an active research topic for over a decade. Over that time, many developers have focused on creating alignment systems and methods to find simple 1-to-1 equivalence matches between two ontologies. However, very few alignment systems focus on finding complex correspondences. There are several reasons for this limitation. First, there are no widely accepted alignment benchmarks that contain such complex relationships. Second, the traditional evaluation metrics like precision, recall, and f-measure are not accurate enough to evaluate the performance of a complex alignment system. And third, the approaches most commonly used to find simple equivalences do not handle the increased computational complexity of finding complex equivalences well. Therefore, it becomes a big challenge for many developers to create and evaluate the systems. In this paper, in order to advance the development of ontology matching, we seek to address the problem by first developing potential complex alignment benchmarks from real-world ontologies. In addition, we utilize traditional automated alignment systems to suggest complex correspondences, and finally plan to achieve our ultimate goal of creating and evaluating our own complex alignment system based on logical RDF data compression.

1 Problem Statement

Similar to database integration, ontology alignment is an important process in enabling computers to query and reason across the many linked open datasets on the semantic web. This is a difficult challenge because the ontologies underlying different linked datasets can vary in terms of subject area coverage, level of abstraction, ontology modeling philosophy, and even language.

In order to solve this problem, an ontology alignment system aims to identify the entity relationships between two and more ontologies. Such relationships have a wide range of complexity, from basic 1-to-1 equivalence to arbitrary m-to-n relationships. However, over a decade, the majority of the studies in the field still focuses on the simplest end of this scale – finding 1-to-1 equivalence relations between ontologies. Very few ontology alignment systems and methods are developed to uncover complex relations. The reasons for this limitation may lie in the following. First, there are no widely used and accepted ontology alignment benchmarks that involve complex relations. Without these benchmarks, even if there were a complex alignment system, it is not only very hard to evaluate if this system could correctly detect the complex correspondences, but also it is a challenge to evaluate whether the system is comprehensive enough to find all different kinds of complex patterns. Second, the traditional approach of precision, recall, and f-measure does not seem fine-grained enough to evaluate complex correspondences. A better version of precision and recall is needed [2].

This work seeks to progress in the direction of fostering the development of research activities in the field of complex ontology alignment. We firstly show that real-world ontologies involve many complex relations. And based on these real-world datasets, we develop high quality complex alignment benchmarks, including creating complex alignments and categorizing them into complex patterns. In addition, to decrease the complexity of detecting complex relations, we leverage automated alignment systems to uncover and suggest possible complex relations. Moreover, we plan to apply logical RDF compression with the results that are generated by traditional automated alignment systems to create a new complex alignment system.

2 Relevancy

Ontology alignment seeks to address the conceptual heterogeneities between ontologies. Over a decade, the community of this field remains on creating and improving the algorithms of finding simple alignments. The reason that the researchers do not really dig into the complex alignment for such a long time, is that the community is still discovering and analyzing how to reach the goal. Nowadays, the research related to simple alignment has been well studied. It is actually a good timing to move on to complex ontology alignment, because more and more good alignment systems and algorithms have been published, and also more and more data are populated into ontologies and published as linked open data, the applications that utilize these LODs are required to involve ontology matching and data integration processes [3]. In addition, due to the complexity of the alignments between ontologies, only identifying traditional simple 1-to-1 alignment is not enough to fulfill the growing high demand of most of these applications. Therefore, it is necessary to create complex alignment systems and methods to uncover complex relations in real-world use cases.

This work focuses on addressing the problem from several different perspectives. We prepared benchmarks that involve complex relations from real-world ontologies and will try to distribute them as a new track in the Ontology Alignment Evaluation Initiative (OAEI), which was started in 2005 with the intent to allow researchers in the field to compare the performance of their approaches on a consistent set of benchmarks over time. Since then, it has been more convenient for researchers from different organizations to test their methods. Second, as we've seen, it is a difficult challenge to detect complex relations. This work seeks to narrow down this issue by leveraging traditional alignment systems to suggest possible complex candidates. This would be a valuable starting point for determining the exact relation. Moreover, our ultimate goal in this work is to create a complex alignment system that has better performance than traditional automated alignment systems.

3 Related work

Regarding the creation of complex alignment benchmarks, Thieblin [11] is creating a complex alignment benchmark using the Conference track ontologies within the OAEI. This work is partially completed, and at the time of this writing it covers three of the seven ontologies. In addition, we are collaborating with them (under their direction) to complete the dataset and prepare a new task in OAEI to evaluate complex alignment systems.

However, even though there are no widely accepted and used benchmarks that involve complex relations, some researchers still tried to create alignment systems and evaluate them using their ow manually developed reference alignment. BLOOMS [3] is an alignment system based on Wikipedia to detect the subsumption relations. Other subsumption systems have evaluated the precision of their approach by manually validating relations produced by their system, while foregoing an assessment of recall [9]. There are some more general approaches based on complex patterns to detect complex correspondences. Ritze *et al* [7, 8] proposed several complex correspondences patterns. Such as: Class by Attribute Value, Class by Attribute Type, Class by Inverse Attribute Type, Inverse Properties, and Property Chain. In addition, Ritze also utilized linguistic analysis techniques, like detection of antonymy, active form, etc to help detect these complex patterns. Other similar work was done by Šváb-Zamazal and Svátek [10]. It firstly detected N-ary relations in the source ontology. And then, it matched the detected N-ary relations to an object property in the target ontology.

Our work differs from the above methods in several aspects. First, we focus on real-world ontologies, which we found that these datasets are not only used by academic researchers, but also the industries and governments to develop applications for the usage of normal human life. There are some interesting relations that have not yet been mentioned in the current benchmark from OAEI. In addition, the instance data of these real-world ontologies are ready to be used as additional information to help improve the performance of alignment process. In contrast to this, significant instance data is not readily available for most of the OAEI Conference Track ontologies. Moreover, regarding the creation of a complex alignment, instead of comparing each entity in the source ontology to each entity in the target ontology, we apply logical RDF compression to list a set of available rules, and narrow down them based on the suggestion generated by traditional alignment systems to finally output the complex relation. More details are discussed in Section 5.

4 Research Questions and Hypotheses

The research questions that we plan to address are listed as follows:

1. Do real-world ontology alignments contain complex relations?
2. How well do traditional automated alignment systems work on real-world matching tasks that contain complex relations?
3. Can we create an automated alignment system that performs better than traditional alignment systems on finding complex relations that exist between ontologies?

Our hypotheses associated to the above research questions are the following:

1. Most real-world ontologies contain many complex relations.
2. Traditional automated alignment systems may not be able to identify complex relationships directly, but they may be able to suggest the atomic entities involved in such relations.
3. A complex alignment system that leverages logical RDF compression can effectively identify complex relations between ontologies.

5 Approach

Hypothesis 1 Our previous work with the NSF EarthCube Initiative and the US Geological Survey involved the time consuming task of manually aligning several real-world ontologies. These alignments have been discussed and evaluated by domain experts and ontology engineers to guarantee that they are of high quality. We will inventory these alignments, along with any other real-world alignments we can acquire, to answer the first research question: Do real-world ontology alignments contain complex relations?

Hypothesis 2 To answer the second research question, "How well do traditional automated alignment systems perform on real-world matching tasks that contain complex relations?", we plan to first evaluate several state of the art alignment systems on the alignment tasks mentioned above. Since traditional alignment systems only attempt to identify simple relations between ontologies, their performance will be limited to the percentage of the alignments that involve these types of relations. However, it is possible that these systems, while they cannot identify the precise relationship that holds between an entity in the source ontology and two or more entities in the target ontology, they can at least identify the entities involved in the relationship. For example, in the relation below, the class $Mischaakusaakihiikin$ in Cree ontology is equivalent to the intersection of instances of $LakeOrPond$ and entities that $isContainedBy$ a $SwampOrMarsh$ in the SWO ontology. While a traditional alignment system cannot identify things like intersection or value restrictions, it may be able to determine that $LakeOrPond$, $isContainedBy$, and $SwampOrMarsh$ are related in some way to $Mischaakusaakihiikin$. To check this, for each entity e_s

in source ontology O_s, we will use the automated alignment systems to give a list of candidates e_t in the target ontology O_t, ordered by the similarity assigned to them by the alignment system. We will evaluate the performance against the benchmark using mean reciprocal rank [6].

```
EquivalentClasses(cree:Mischaakusaakihiikin
        ObjectIntersectionOf(swo:LakeOrPond
                ObjectSomeValuesFrom(swo:isContainedBy
                        swo:SwampOrMarsh)))
```

Hypothesis 3 As we mentioned, the ultimate goal of this work is to see if we can create an automated alignment system that effectively identifies complex relationships that exist between two ontologies. Our planned approach is to create an extensional matcher (i.e. one that relies upon instance data) that leverages logical RDF compression [5]. Logical RDF compression uses the FP-Growth data mining algorithm to generate rules that can be stored in lieu of the triples they are based on. For example, say that a linked dataset contains triples about university students. There might be many triples of the form <ind1 hasMajor ComputerScience> and many corresponding triples of the form <ind1 isEnrolledIn CollegeOfEngineering> because, according to this dataset, all Computer Science majors are enrolled in the College of Engineering. Logical RDF compression would replace the second set of triples with a single rule: if x is hasMajor ComputerScience then x isEnrolledIn CollegeOfEngineering, and these triples could then be generated on-the-fly in response to queries, thereby saving space in the linked dataset. While logical RDF compression seeks to find any rules that can be used to shrink the dataset, it is possible that some of these rules represent meaningful semantic relations that hold between entities. For example, if hasMajor ComputerScience exists in one ontology and isEnrolledIn CollegeOfEngineering exists in another ontology, then it may be possible to infer the relation below.

```
SubClassOf(ObjectSomeValuesFrom(ont1:hasMajor ont1:ComputerScience)
    ObjectSomeValuesFrom(ont2:isEnrolledIn ont2:CollegeOfEngineering))
```

Because the FP-Growth algorithm underlying logical RDF compression can generate a very large number of rules, some mechanism must be put in place to choose the more semantically meaningful rules rather than the ones that result in the most compression. Our planned approach for this is to choose rules that involve the entities suggested by traditional alignment systems.

The overall work flow is shown in Figure 1. We first apply the traditional alignment systems to suggest the candidates as we described above. And then, we use RDF compression [5] on the source ontology to list a set of compression rules. Based on the suggested candidates from traditional alignment systems, we can create a filter to pick up the compression rules, and finally output the complex relations.

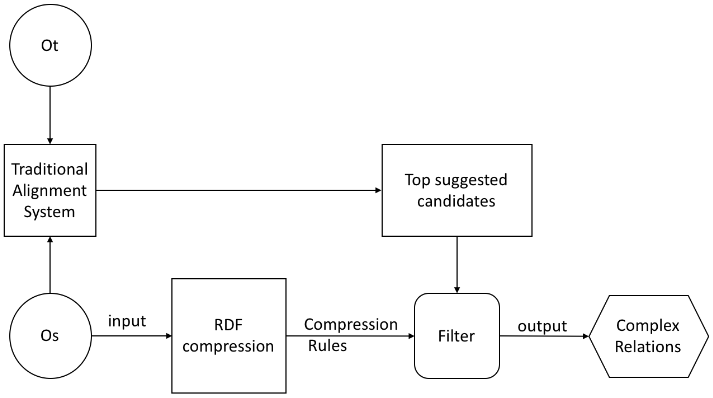

Fig. 1. The Work Flow of the Proposed Complex Alignment System

6 Preliminary Results

We have some preliminary results in terms of finding complex relations from real-world ontologies. There are two different datasets that we are currently working on. One dataset is from GeoLink project[1] that was funded under the U.S. National Science Foundation's EarthCube initiative. Another dataset is a set of ontologies from surface water domain. Based on these two sets of ontologies, we have developed the alignments in consultation with domain experts from different institutions. In GeoLink datasets, the number of classes, object properties, and data properties in GeoLink base ontology (GBO) and GeoLink modular ontology (GMO) are showed in Table 1. We found that the 87 out of 111 relations in them are complex relations, including not only class subsumption, property subsumption, property chain equivalence, and property chain subsumption that were introduced in [7], but also some typecasting relations. The idea of typecasting, and why it is important in ontology modeling, is formally introduced and discussed in [4]. Moreover, the alignment is also available in both EDOAL and rules syntax for the purpose of manipulating and reading respectively. The full dataset has been uploaded to the FigShare[2]. We also wrote a paper to describe the creation and submitted to ISWC2018, which is currently under-review.

In hydrography dataset, it consists of four ontologies. (a) The Surface Water Ontology (SWO), which was originally presented in [12], was developed by the US Geological Survey (USGS). (b) The Hydro3 ontology was developed by individuals at the University of Maine in order to support expanded gazetteer functions using topology and semantic inference [13]. (c) The HydrOntology is a non-English ontology, which was developed by the Spanish National Geographic Institute (IGN) [1]. (d) Cree surface water ontology is in a language, Cree, which

[1] https://www.geolink.org/
[2] https://doi.org/10.6084/m9.figshare.5907172

Table 1. The Number of Classes, Object Properties, and Data Properties in Both GeoLink Ontologies

Ontology	Classes	Object Properties	Data Properties
GBO	40	149	49
GMO	156	124	46

is spoken by some of the native inhabitants of northern Canada. The reason of choosing these four ontologies is that Hydro3, HydrOntology and Cree have a large degree of overlap with SWO. Therefore, we utilize these four ontologies to create a reference alignment manually. Table 2 shows the number of classes, and object properties, and data properties. And we found that the 84 out of 197 relations in them are complex relations. We are currently writing a paper about evaluating the performance of traditional automated alignment systems on this dataset. The traditional automated alignment systems are able to find the simple relationships. But, they may not be able to identify these complex relationships directly. We hypothesize that they may be able to suggest some possible complex relations partially. Moreover, it is able to greatly narrow down the entities and mitigates the high complexity of computation.

Table 2. The Number of Classes, Object Properties, and Data Properties in Hydrography Ontologies

Ontology	Classes	Object Properties	Data Properties
SWO	85	20	1
Hydro3	22	34	0
HydrOntology	154	47	75
Cree	83	21	7

7 Evaluation Plan

In this section, we introduce the evaluation plan for each research question. For research question 1, as we showed in Section 6, it is considered successful that we have found many complex alignments in real-world ontologies. In addition, we are also preparing to incorporate the dataset into OAEI as a new track for other researchers accessing it. For research question 2, after achieving the list of entities involved in a complex relation, we will evaluate the performance against the benchmark using mean reciprocal rank as we discussed in Section 5. For research question 3, it is a challenge to evaluate the performance of a complex alignment system. The traditional precision, recall, and f-measure metrics do not seem fine-grained enough. For example, there is a relation between Hydro3 and SWO, if one alignment system identified this:

```
EquivalentClasses(
        ObjectIntersectionOf(
                hydro3:Hydrographic_Feature
                hydro3:Hydrographic_Structure
                hydro3:Boundary)
        swo:HydrographicFeature))
```

and another identified this:

```
SubClassOf(
        ObjectUnionOf(
                hydro3:Hydrographic_Feature
                hydro3:Island
                hydro3:Shore)
        swo:HydrographicFeature))
```

Based on the reference alignment, we need a metric to consider the first system "more correct" than the second. We plan to develop a performance metric that more accurately reflects the performance of a complex alignment system. Another challenge is that, to the best of our knowledge, there are no existing complex alignment systems against which to compare our approach. Therefore, we might consider evaluating the performance based on our manually created reference alignment.

8 Reflections

It is primarily difficult to identify complex relationships between ontologies because of computational complexity. A naive approach would need to compare every entity in the source ontology to every possible *combination* of entities in the target ontology, which is not feasible. Instead of doing this, our proposed approach has a good chance of success because it is based on a logical RDF compression method that has already been shown to be applicable to large datasets, and we also can further limit the search space by using the output from traditional alignment systems to narrow the focus. There are some reflections. The performance of using logical RDF compression in our alignment system is primarily based on the *Abox* information in the ontology. It is still not clear that how to apply our alignment algorithm to a more generalized scenario. However, our approach is feasible, and can be a good starting point to achieve the ultimate goal in the future.

Acknowledgments I am very grateful to my advisor Dr. Pascal Hitzler and Dr. Michelle Cheatham for their support and constructive suggestions of this research proposal.

References

1. Blázquez, L.M.V., Gargantilla, J.Á.R., López-Pellicer, F.J., Corcho, Ó., Nogueras-Iso, J.: An approach to comparing different ontologies in the context of hydrographical information. In: Information Fusion and Geographic Information Systems, Proceedings of the Fourth International Workshop, IF&GIS 2009, 17-20 May 2009, St. Petersburg, Russia. pp. 193–207 (2009)
2. Euzenat, J.: Semantic precision and recall for ontology alignment evaluation. In: IJCAI 2007, Proceedings of the 20th International Joint Conference on Artificial Intelligence, Hyderabad, India, January 6-12, 2007. pp. 348–353 (2007)
3. Jain, P., Hitzler, P., Sheth, A.P., Verma, K., Yeh, P.Z.: Ontology alignment for linked open data. In: The Semantic Web - ISWC 2010 - 9th International Semantic Web Conference, ISWC 2010, Shanghai, China, November 7-11, 2010, Revised Selected Papers, Part I. pp. 402–417 (2010)
4. Krisnadhi, A.A., Hitzler, P., Janowicz, K.: On the capabilities and limitations of OWL regarding typecasting and ontology design pattern views. In: Ontology Engineering - 12th International Experiences and Directions Workshop on OWL, OWLED 2015, co-located with ISWC 2015, Bethlehem, PA, USA, October 9-10, 2015, Revised Selected Papers. pp. 105–116 (2015)
5. Pan, J.Z., Gómez-Pérez, J.M., Ren, Y., Wu, H., Wang, H., Zhu, M.: Graph pattern based RDF data compression. In: Semantic Technology - 4th Joint International Conference, JIST 2014, Chiang Mai, Thailand, November 9-11, 2014. Revised Selected Papers. pp. 239–256 (2014)
6. Radev, D.R., Qi, H., Wu, H., Fan, W.: Evaluating web-based question answering systems. In: Proceedings of the Third International Conference on Language Resources and Evaluation, LREC 2002, May 29-31, 2002, Las Palmas, Canary Islands, Spain (2002)
7. Ritze, D., Meilicke, C., Sváb-Zamazal, O., Stuckenschmidt, H.: A pattern-based ontology matching approach for detecting complex correspondences. In: Proceedings of the 4th International Workshop on Ontology Matching (OM-2009) collocated with the 8th International Semantic Web Conference (ISWC-2009) Chantilly, USA, October 25, 2009 (2009)
8. Ritze, D., Völker, J., Meilicke, C., Sváb-Zamazal, O.: Linguistic analysis for complex ontology matching. In: Proceedings of the 5th International Workshop on Ontology Matching (OM-2010), Shanghai, China, November 7, 2010 (2010)
9. Suchanek, F.M., Abiteboul, S., Senellart, P.: Paris: Probabilistic alignment of relations, instances, and schema. Proceedings of the VLDB Endowment **5**(3), 157–168 (2011)
10. Šváb-Zamazal, O., Svátek, V.: Towards ontology matching via pattern-based detection of semantic structures in owl ontologies. In: Proceedings of the Znalosti Czecho-Slovak Knowledge Technology conference (2009)
11. Thieblin, E., Haemmerle, O., Hernandez, N., Trojahn, C.: Towards a complex alignment evaluation dataset. In: Proceedings of the 16th International Conference on Ontology Matching-Volume 2032. CEUR-WS. org (2017)
12. Varanka, D.E., Usery, E.L.: An applied ontology for semantics associated with surface water features. Land Use and Land Cover Semantics: Principles, Best Practices, and Prospects p. 145 (2015)
13. Vijayasankaran, N.: Enhanced place name search using semantic gazetteers (2015)