Identifying Complex Semantic Matches

Brian Walshe

KDEG & FAME, Department of Computer Science and Statistics,
Trinity College Dublin
walshebr@scss.tcd.ie

Abstract. One-to-one correspondences are not always sufficient to accurately align ontologies, and instead complex correspondences with conditions and transformations may be required. Correspondence patterns provide models which can be used to guide the process of developing complex correspondences. However, it is necessary to first identify which pattern to apply to a given alignment problem. This PhD proposes the development of algorithms, methods and processes for refining elementary correspondences between concepts or relations into complex ones by identifying which correspondence pattern best represents a given correspondence. To date an evaluation of a system to refine correspondences between classes in the YAGO and DBpedia ontologies has been completed. This evaluation showed that for a subsumption correspondence, a training set of 30 instances of the class being mapped was sufficient to refine the match to a conditional one 89% of the time. Hence we have shown that this is a promising approach for correspondences with a conditional element, and correspondences with a translation element will be examined next.

1 Motivation

When organizations wish to work together, integrating their data is usually a significant challenge. As each organization's data resources have often been developed independently, they are subject to heterogeneity – primarily semantic, syntactic and structural. The use of ontologies has long been identified as an effective way of facilitating the integration process, seeing use in diverse fields, including for example, biomedicine and high-energy physics [1].

Suitable techniques for discovering and describing matches between ontologies are still very much an open problem [2]. Matching ontologies is a demanding and error prone task. Not only does it require expert knowledge of the matching process, but also an in-depth understanding of the subject the ontologies describe, and typically knowledge of the principals used to construct the ontologies. The ontologies can differ in scope, granularity and coverage, they can use different paradigms to represent similar concepts, or use different modeling conventions [3]. Therefore producing high quality correspondences is typically semi-automated – an expert user approves and refines match candidates produced by an automatic semantic matcher tool [4].

Semantic matcher tools typically focus on detecting schema-level equivalence relations. However elementary correspondences between named ontology elements are often not sufficient for tasks such as query rewriting, instance translation, or instance

E. Simperl et al. (Eds.): ESWC 2012, LNCS 7295, pp. 849–853, 2012.

mediation. Complex correspondences which contain conditions and transformations are necessary to satisfy these real-world use cases. Recent semantic matching research has produced a taxonomy of complex correspondences, called Correspondence Patterns by Scharffe [5]. As of version 4.0, the INRIA Alignment API [6] has introduced the Expressive Declarative Ontology Alignment Language (EDOAL) for expressing complex relationships of this nature, but support among semantic matchers for discovering matches that make full use of EDOAL is still limited.

While progress has been made in matching research, the potential use cases for semantic matching have also expanded with the realization of a web of data as Linked Data. Linked Open Data (LOD) represents a large amount of the content contained in the current Semantic Web, with DBpedia [7] currently describing 3.64 million things, 1.83 million of which are classified in a consistent ontology[1]. As these sources overlap and complement each other, mapping between them is important. However the shallow taxonomies typically used in LOD sources means that complex alignments are required if we wish to use this instance data with the richer structure found in a richer ontology. For example an instance of type *AmericanFilmDirector* in YAGO must become an instance of type *Person*, with the *occupation* attribute set to *Director* and *nationality* set to *American* if it is to be understood in terms of the DBpedia schema.

The research question for this PhD is as follows: *To what extent can elementary semantic matches be refined to complex matches – containing constraints and transformations – using a semi-automatic process based on classifying the matches against the Correspondence Patterns scheme.* In this context *elementary* matches are ones which match named ontology elements using relations such as subsumption. Existing semantic matching tools rely primarily on analyzing the structural information in ontologies, however as many data sources on the Semantic Web contain large amounts of instance data, this work will focus on techniques which use instance data to discover relationships in the structural data.

The following section of this paper outlines the state of the art in semantic matching and related techniques, as well as some of the shortcomings of current solutions. Section 3 discusses the proposed research approach, and the research methodology that will be employed. Section 4 outlines some preliminary results from an initial experiment.

2 State of the Art

This section outlines a survey of the state of the art in the semantic matching process, including tools, processes and formats for describing matches. It also includes a summary of schemes for classifying forms of heterogeneity, as well as machine learning tools which will be of benefit in identifying complex matches.

While it is held as impossible to completely automate the alignment process [4], several tools have been developed to assist a human operator with the endeavor. These include both automated candidate match generation [8], and graphical interfaces to assist with discovering and confirming correspondences between ontologies [9].

[1] http://blog.dbpedia.org/2011/09/11/dbpedia-37-released-including-15-localized-editions/

The INRIA Ontology Alignment API [6] provides a comprehensive set of automated match generators, drawn from the methods described by Shvaiko and Euzenat [8]. It also provides a format for describing matches, so that they may be exchanged.

Current semantic matching tools focus on discovering *equivalence* (≡) relationships, and less commonly *less general* (⊑), *more general* (⊒) and *disjointness* (≢) relationships. Schemes exist for classifying more detailed forms of relationships, these include Correspondence Patterns [5], and the THALIA framework [10]. The EDOAL [6] language provides a method for describing more complicated correspondences, using an OWL-like syntax, and was developed in conjunction with correspondence patterns. As yet, there are few matchers capable of generating the complex matches that EDOAL allows [6]. Ritze et. al. [11] describe a first attempt process for detecting complex matches, and Sváb-Zamazal et. al. [12] provide a set of pattern based tools for describing and managing these matches by relating them to ontology design principals.

The field of Machine Learning provides many tools which could be of use in analyzing matches between the less structured elements found in Linked Data. The Weka suite [13] provides a range of these tools as both an API and as part of a GUI. These include attribute selection tools such as Information Gain, and regression analysis tools.

3 Research Approach and Methodology

The focus of this research will be on the development of algorithms, methods and processes for refining elementary semantic relationships such as equivalence into more complex correspondences with conditions and transformations. The novelty of this approach lies in the use of instance element information to identify a correspondence pattern that best applies to a given ABox relationship and then using the pattern to guide the production of a complex correspondence.

For example, the *Class by Attribute Value* correspondence pattern occurs when a class in one ontology is equivalent to a class in a second ontology defined by instances with a specific property value. This pattern can be detected by the use of attribute selection methods, and once detected a declarative alignment may be created specifying the classes that match and the attribute and value condition necessary for the match to hold true. Section 4 outlines an experiment where an Information Gain based attribute selection method was used to detect correspondences of type *Class by Attribute Value* and *Class by Attribute Existence*.

A further pattern is that of the *Property – Relation Correspondence*, where a property in one ontology corresponds with a relation in another ontology. Here clustering could be used to group data values in the range of the property so they may be matched to the individuals in the range of the relationship. This would allow the creation of a declarative match consisting of a metric and cutoff point matching to a named individual. *Property Value Transformations* are another form of correspondence that can occur. This pattern would be detected when the values of matched properties for matched instances differ in some consistent way. For numerical properties, linear regression could be used to investigate what transformation is occurring, for example unit conversions or combinations such as o_1#price+o_1#tax → o_2#price.

Many of the techniques that will be investigated are reliant on analyzing instance data contained in the ontologies, and one possible limitation is that if the ontologies being mapped do not have suitable instance data these techniques will not be applicable. Because of this there will be a focus on Linked Data, as Linked Data sources contain large amounts of instance data, and often overlap to some extent. A further obstacle to this research will be the ability to measure the effectiveness of the matching techniques against a standard test set such as those used by the OAEI [14].

The methodology of this PhD consists of a literature review to establish the current state of the art in semantic matching tools, methods for their evaluation, and formats available for describing alignments. Following from this, techniques for detecting restriction type Correspondence Patterns such as Class by Attribute Value and Class by Attribute Existence and converting them to complex alignments will be developed. Techniques for analyzing Property Value Transformation correspondence patterns by regression or other means shall also be developed. A suitable test suite to evaluate the these techniques will be required, which should be made public to allow comparison with other matchers capable of producing complex matches

To date a tool has been developed which is capable of refining class matches to include declarative restrictions. An experiment was carried out to evaluate the ability of this system, and the results are described in the following section.

4 Current Work

An evaluation was carried out of the ability of a system to refine elementary correspondences between the class *Person* in DBpedia [7] and several classes of occupation in YAGO [15] to produce complex correspondences. Using matched instances of these classes that had been identified in both sources, attribute selection was used to test if restriction type matches were occurring between the classes, and determine which attribute these restrictions were dependent on. This evaluation demonstrated that the Information Gain measure is a suitable scoring function for finding the attribute and attribute value to condition matches of type *Class by Attribute Value* and *Class by Attribute Existence*. This Information Gain function allowed us to reliably select a gold standard correspondence pattern – consisting of an attribute and value to condition – as our top result in two of the four mappings tested, and reliably returns the best correspondence pattern in the top five results for all four test cases. In the cases where the search algorithm did not return the best correspondence pattern as the top result, this was because there were several patterns that could be considered valid and selecting the "best" among these was difficult. The IG score requires that a suitable training set of instances be used, and the evaluation demonstrated that a random sample of 30 instances was sufficient to rank the gold standard attribute and attribute-value pair in the top 5 results at least 89% of the time. The results of this experiment have been accepted for publication at the DANMS 2012 workshop.

Acknowledgement. This research is supported by the Science Foundation Ireland (Grant 08/SRC/I1403) as part of the FAME Strategic Research Cluster (www.fame.ie).

References

1. Weidman, S., Arrison, T.: Steps toward large-scale data integration in the sciences: summary of a workshop. National Academy of Sciences, 13 (2010)
2. Shvaiko, P., Euzenat, J.: Ontology matching: state of the art and future challenges. IEEE Transactions on Knowledge and Data Engineering (preprint, 2012)
3. Klein, M.: Combining and relating ontologies: an analysis of problems and solutions. In: Proc. of Workshop on Ontologies and Information Sharing, IJCAI, pp. 53–62 (2001)
4. O'Sullivan, D.: The OISIN framework: ontology interoperability in support of semantic interoperability. PhD thesis, Trinity College Dublin (2006)
5. Scharffe, F.: Correspondence patterns representation. PhD thesis, University of Innsbruck (2009)
6. David, J., Euzenat, J., Scharffe, F., Trojahn dos Santos, C.: The alignment API 4.0. Semantic Web 2(1), 3–10 (2011)
7. Bizer, C., et al.: DBpedia - A crystallization point for the Web of Data. Web Semantics: Science, Services and Agents on the World Wide Web (7), 154–165 (2009)
8. Shvaiko, P., Euzenat, J.: A Survey of Schema-Based Matching Approaches. In: Spaccapietra, S. (ed.) Journal on Data Semantics IV. LNCS, vol. 3730, pp. 146–171. Springer, Heidelberg (2005)
9. Falconer, S.M., Storey, M.-A.: A Cognitive Support Framework for Ontology Mapping. In: Aberer, K., Choi, K.-S., Noy, N., Allemang, D., Lee, K.-I., Nixon, L.J.B., Golbeck, J., Mika, P., Maynard, D., Mizoguchi, R., Schreiber, G., Cudré-Mauroux, P. (eds.) ASWC 2007 and ISWC 2007. LNCS, vol. 4825, pp. 114–127. Springer, Heidelberg (2007)
10. Hammer, J., Stonebraker, M., Topsakal, O.: THALIA: Test Harness for the Assessment of Legacy Information Integration Approaches. In: Proc. of the Int. Conference on Data Engineering (ICDE), pp. 485–486 (2005)
11. Ritze, D., Meilicke, C., Sváb-Zamazal, O., Stuckenschmidt, H.: A pattern-based ontology matching approach for detecting complex correspondences. In: Proc. of Int. Workshop on Ontology Matching, OM (2009)
12. Sváb-Zamazal, O., Daga, E., Dudás, M.: Tools for Pattern-Based Transformation of OWL Ontologies. In: Proc. of Int. Semantic Web Conference (2011)
13. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: an update. ACM SIGKDD Explorations Newsletter 11(1), 10–18 (2009)
14. Euzenat, J., et al.: First results of the Ontology Alignment Evaluation Initiative 2011. In: Proc. of 6th Int. Workshop on Ontology Matching, OM 2011 (2011)
15. Suchanek, F.M., Kasneci, G., Weikum, G.: Yago: A large ontology from wikipedia and wordnet. Web Semantics: Science, Services and Agents on the World Wide Web 6(3), 203–217 (2008)