Dealing with Uncertainty Issues in Complex Ontology Matching

Ying Wang[1] and **Weiru Liu**[1] and **David Bell** [1]

Abstract. Ontology mapping is one of the most important tasks for ontology interoperability and its main aim is to find semantic relationships between entities of two ontologies. However, most of the current techniques suffer from some kind of drawbacks as listed below: (a) most of them only consider 1:1 mappings; (b) most of them do not consider the importance of uncertainty in ontology mapping. In this paper we consider the following two issues that have been the focus of our ongoing research: (a) how to produce complex mappings (m:1 or 1:m and m:n) and (b) how to deal with uncertainties in the process of ontology mapping.

1 INTRODUCTION TO THE PROBLEM

Research and development on ontology mapping (or matching) has attracted huge interests and many mapping methods have been proposed. Comprehensive surveys on recent developments of ontology mapping can be found in [10, 11].

Considerable efforts have been devoted to implement ontology mapping systems, especially 1:1 mappings. However, complex mappings are also pervasive and important in real world applications. In [10], an example was given to illustrate the importance of complex mappings in schema mapping research. We think that the same issue exists in ontology mapping and the example is applicable to ontology mapping. Let us take a look of the example. Given two ontologies O_A and O_B, they contain different entities respectively: **Book** and **Publisher** in O_A; **Title** and **Name** in O_B. It is clear that entities {**Book, Publisher**} of O_A should be matched to {**Title, Name**} of O_B.

Another aspect is that most of the earlier works in this area did not consider uncertainty or imprecision occurred during a mapping, however, in most cases, the mappings between entities produced are imprecise and uncertain. For instance, most automatic ontology mapping tools use heuristics or machine-learning techniques, which are imprecise by their very nature. Even experts are sometimes unsure about the exact matches between concepts and typically assign some certainty ratings to a match [2]. So a matching result is often associated with a weight which can express how close the two entities are as a match. The needs to consider uncertainty in a mapping began to emerge in a number of papers (e.g., [8, 1, 9, 4, 13]) in which Dempster Shafer theory, Bayesian Networks, and rough sets theory are used to deal with different aspects of mapping or ontology descriptions (e.g., concept subsumptions).

The rest of the paper is organized as follows. Section 2 presents the set-inclusion based approach we proposed for dealing with complex matching. Section 3 describes the clustering-based approach we developed for handling uncertainties in ontology mapping. Section 4 concludes the paper.

[1] School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, Belfast, BT7 1NN, UK , email: {ywang14, w.liu, da.bell}@qub.ac.uk

2 A SET INCLUSION BASED ONTOLOGY MAPPING APPROACH

Before we introduce this new ontology mapping approach, we first describe a new method to represent entities in ontologies. Traditionally, the concept names of entities are used directly in mapping. This representation method does not consider the hidden relationships between concept names of entities, so it cannot reflect the complete meaning of the concept names of entities. Here we explore a new representation method for entities. For the multi-hierarchical structure of ontology, we observe that for each concept in this concept hierarchy, its complete meaning is described by a set of concept names. In other words, there is a kind of *inclusion relationship* among these concepts. So for any concept name of entity C in an ontology, we can represent it by a new method as follows. First, we find the branch which has the concept C. Second, we collect those concepts along the path between C and the root node to form a set. We use this new set to represent C.

Once each entity is represented by a set of words, we compute the similarities between entities. Here, we choose the *Linguistic-based matcher* (which uses domain specific thesauri to match words) and the *Structure-based matcher* (which uses concept-hierarchy theory) to compute similarities (we utilize Linguistic-based matcher because the performance of this matcher is good for similar or dissimilar words. Please refer to [12] for details).

As a result, we obtain a set S_1 consisting of mapping candidates such that from each entity in ontology O_1, a similarity value is obtained for every entity in ontology O_2. Following this, we select the best mapping entity in O_2 for each entity in O_1 and these best mapping results constitute another set S_2. In S_2, we search all the mapping results to see if there exist multiple source entities in O_1 that are mapped to the same target entity in O_2. If so, we apply a new algorithm based on Apriori algorithm [3] to decide how many source entities in O_1 should be combined together to map onto the same entity in O_2.

We use the OAEI 2007 Benchmark Tests and we now compare the outputs from our system (denoted as SIM) to the results obtained from *ASMOV, DSSim, TaxoMap* and *OntoDNA* algorithms which were used in the 2007 Ontology Alignment Contest [2] in which almost all the benchmark tests describe Bibliographic references and the details are given in Table 1. In Table 1, p for precision, r for recall, f for f-measure. Our study shows that this method significantly improves the matching results as illustrated in our experiments.

[2] http://oaei.ontologymatching.org/2007/results/

Table 1. Comparison of Experiment Results

Datasets	SIM			ASMOV			DSSim			TaxoMap			OntoDNA		
	p	r	f	p	r	f	p	r	f	p	r	f	p	r	f
101	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
103	100	100	100	100	100	100	100	100	100	100	34	51	94	100	97
104	100	100	100	100	100	100	100	100	100	100	34	51	94	100	97
203	100	100	100	100	100	100	100	100	100	NaN	0.00	NaN	94	100	97
204	86	84	85	100	100	100	96	91	93	92	24	38	93	84	88
205	47	44	46	100	100	100	94	33	49	77	10	18	57	12	20
208	86	83	85	100	100	100	95	90	92	NaN	0	NaN	93	84	88
209	49	41	45	92	90	91	91	32	47	NaN	0	NaN	57	12	20
221	82	82	82	100	100	100	100	100	100	100	34	51	93	76	83
222	89	92	91	100	100	100	100	100	100	100	31	47	94	100	97
224	100	100	100	100	100	100	100	100	100	100	34	51	94	100	97
225	100	100	100	100	100	100	100	100	100	100	34	51	94	100	97
228	100	100	100	100	100	100	100	100	100	100	100	100	53	27	36
230	73	90	81	99	100	99	97	100	98	100	35	52	91	100	95
231	100	100	100	100	100	100	100	100	100	100	34	51	94	100	97
232	82	82	82	100	100	100	100	100	100	100	34	51	93	76	84
233	52	52	52	100	100	100	100	100	100	100	100	100	53	27	32
236	100	100	100	100	100	100	100	100	100	100	100	100	53	27	32
237	93	97	95	100	100	100	100	100	100	100	31	47	94	100	97
239	88	100	94	97	100	98	97	100	98	100	100	100	50	31	38
241	58	58	58	100	100	100	100	100	100	100	100	100	53	27	32
246	88	100	94	97	100	98	97	100	98	100	100	100	50	31	38
301	43	45	44	93	82	87	82	30	44	100	21	35	88	69	77
302	34	53	42	68	58	63	85	60	70	100	21	35	90	40	55
304	51	49	50	95	96	95	96	92	94	93	34	50	92	88	90

Overall, we believe that the experimental results of our system are good. Although on individual pair of ontologies, our results are less ideal than the *ASMOV* system and *DSSim*, however, our results are better than *TaxoMap* system and *OntoDNA* system on most pairs of matching. The performances of these three different approaches, i.e., *ASMOV, DSSim* and our system *SIM* are good for almost the whole data set from Test 101 to Test 246, but our system does not perform well for Test 205, Test 209, Test 233 and Test 241. The performance of all these five systems are not very good for the data set from Test 301 to Test 304. Below we analyze the reasons for this.

For Test 101 vs 103 and vs 104, the two ontologies to be matched contain classes and properties with exactly the same names and structures, so every system that deploys the computation of similarities of names of entities can get good results. Test 201-210 describe the same kind of information as other ontologies, i.e. publications, however, the class names in them are very different from those in the reference ontology Test 101, especially Test 205 and 209, so our system does not obtain good results. The structure of Test 221-247 have been changed although the linguistic features have been maintained, the performance of our system has been affected. Our method is based on the hierarchical structure of an ontology, but for Test 233 and Test 241, these two ontologies have only one layer. When computing the similarity between two concepts in Test 233 and Test 101, such as **MastersThesis** in Test 233 and **MastersThesis** in Test 101. First, our method extends **MastersThesis**. Test 233 only has one layer, so **MastersThesis** can not be changed. Test 101 has three layers, so **MastersThesis** is extended to {**MastersThesis, Academic, Reference**}. The similarity value is reduced and does not reflect the true similarity between these two concepts.

Test 301-304 are real-life BibTeX ontologies which also include different words compared to Test 101 describing publications so the results are similar to Test 205, so we do not get good similarity results from this data set. However we still find some complex mappings (m:1) by using our algorithm to discover the best mapping results, such as for Test 302 vs Test 101, we get {**Collection, Monograph, Book**} mapping to **Book**.

3 CLUSTERING-BASED APPROACH TO COMBINING UNCERTAIN OUTPUTS FROM MULTIPLE ONTOLOGY MATCHERS

We propose a clustering-based approach to combining outputs from multiple ontology matchers (**CCM**). We consider complex mappings between two ontologies O_1 and O_2 which are encoded in OWL. First, we partition entities in ontology O_1 based on *average-linkage clustering algorithm*. This algorithm uses similarity values between entities to do the partitioning, the similarity values are obtained by integrating Lin's matcher [7] (which uses domain specific thesauri to match words) and structure-based method to compute the similarities between entities of O_1 (we utilize Lin's matcher because the performance of this matcher is good for similar or dissimilar words. Please refer to [12] for details). As a result, the similar entities in O_1 are clustered together. Second, for each entity e_{2j} in ontology O_2, we try to find the most appropriate cluster C_{1i} in the collection of clusters created from ontology O_1. Cluster C_{1i} is regarded as the most appropriate for e_{2j} if the similarity value between e_{2j} and C_{1i} is the largest. Third, we deploy four different matchers to calculate the similarity values between a cluster from O_1 and an entity in O_2. We choose several matchers because one matcher analyzes only some aspects of the hypothetical relation between two terms and may lack or omit important information about the relationship between entities [1]. Therefore, if we use more than one matcher, these matchers can complement each other and capture more features about the relationship between entities. Finally, since each match gives a mapping that is not absolutely certain, we apply Dempster-Shafer theory to combine the matching outputs from these four matchers.

We choose two pairs of ontologies, one is **Test 101-205** and another is **russia12**[3] that describe tourism information of Russia. These ontologies are well-known for ontology alignment tests. Their sizes are moderate. To evaluate the mapping quality, here we employ the metric of *correctness* and *f-measure*.

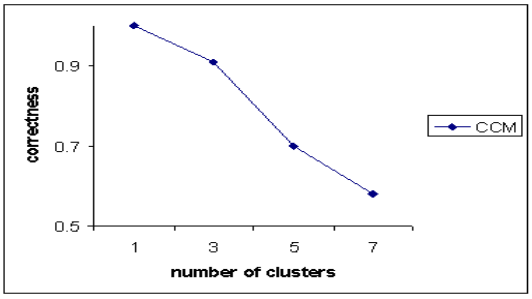

Figure 1. Test 101-205

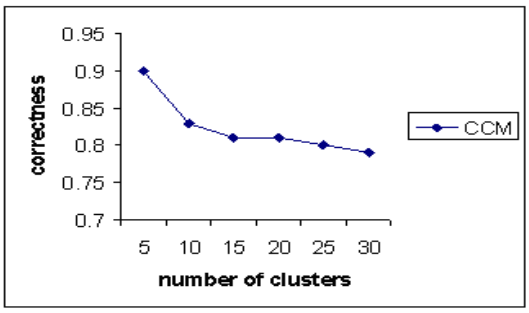

Figure 2. russia12

Both of Figure 1 and Figure 2 show the variation of correctness along with the number of the clusters in cluster mappings. From these two figures, we can see that when the number of clusters increases, the correctness of the cluster mapping decreases. For Figure 1, the names and structures of entities in these two ontologies are very different, so the downward trend of curved line is quick. For Figure 2, the names and structures of entities in these two ontologies are very similar, so the downward trend of CCM is slow when the number of clusters increases.

Table 2. Comparison of Experiment Results

Datasets	approach	number	correctness	f-measure
russia12	CCM	13	0.82	0.30
russia12	BMO	13	0.84	0.56
russia12	PBM	13	0.57	0.65

In Table 2, the comparison results of ontology mapping quality

[3] http://www.aifb.uni-karlsruhe.de/WBS/meh/foam/

and the partitioning quality of CCM, BMO [5] and PBM [6] are presented. The number of cluster mappings is 13. Overall, we believe the DS combination rule is effective although the *f-measure* of CCM is not very good. One of the reasons is that although we utilize a combination method which combine Lin-based matcher and structure method together to compute the similarity between entities in O_1, the results of similarity are still not very good. Another reason is that the matchers we used are based on linguistic features of entities of ontology and they can only handle the problem of mapping from one aspect, meanwhile the similarity results obtained from internal matchers are not very accurate, so when we combine these results by the DS combination rule, some useful but different results are left out.

4 CONCLUSION

Ontology mapping is a difficult task. So for our future work, on the one hand, we will continue improving our current proposed approaches, especially for complex mapping and how to deal with inconsistency produced by different matchers. On the other hand, we will continue investigating the uncertainty issue in ontology mapping and consider how to use different uncertainty theories to deal with different situations in ontology mapping.

REFERENCES

[1] P. Besana., 'A framework for combining ontology and schema matchers with dempster shafer', in *the International Workshop on Ontology Matching (OM'06), collocated with the 5th International Semantic Web Conference (ISWC'06)*, pp. 196–200, (2006).

[2] N. Choi, I.Y. Song, and H. Han, 'A survey on ontology mapping', *In SIGMOD Record*, **35**, 34–41, (2006).

[3] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, 2000.

[4] M. Holi and E. Hyvoen, 'Modeling degrees of conceptual overlap in semantic web ontologies', in *the International Workshop on Uncertainty Reasoning for the Semantic Web (URSW'05), collocated with the the 4th International Semantic Web Conference (ISWC'05)*, pp. 98–99, (2005).

[5] W. Hu and Y. Qu, 'Block matching for ontologies', in *the Proceedings of the 5th International Semantic Web Conference (ISWC'06)*, pp. 300–313, (2006).

[6] W. Hu, y. Zhao, and y. Qu, 'Partition-based block matching of large class hierarchies', in *the Proceedings of the 1st Asian Semantic Web Conference (ASWC'06)*, pp. 72–83, (2006).

[7] D. Lin, 'An information-theoretic definition of similarity', in *the Proceedings of the 15th International Conference on Machine Learning (ICML'98)*, pp. 296–304, (1998).

[8] M. Nagy, M. Vargas-Vera, and E. Motta, 'Dssim-ontology mapping with uncertainty', in *the International Workshop on Ontology Matching (OM'06), collocated with the 5th International Semantic Web Conference (ISWC'06)*, (2006).

[9] R. Pan, Z. Ding, Y. Yu, and Y. Peng, 'A bayesian network approach to ontology mapping', in *the Proceedings of the 4th International Semantic Web Conference (ISWC'05)*, pp. 563–577, (2005).

[10] E. Rahm and P.A. Bernstein, 'A survey of approaches to automatic schema matching', *Journal of VLDB*, **10**, 334–350, (2001).

[11] P. Shvaiko and J. Euzenat, 'A survey of schema-based matching approaches', *Journal of Data Semantics*, **4**, 146–171, (2005).

[12] Y. Wang, W. Liu, and D. Bell, 'Combining uncertain outputs from multiple ontology matchers', in *the Proceedings of the 1st International Conference on Scalable Uncertainty Management (SUM'07)*, pp. 201–214, (2007).

[13] Y. Zhao, X. Wang, and W.A. Halang, 'Ontology mapping based on rough formal concept analysis', in *the Proceedings of the Advanced International Conference on Telecommunications and International Conference on Internet and Web Applications and Services (AICT/ICIW'06)*, p. 180, (2006).