PRIOR System: Results for OAEI 2006

Ming Mao, Yefei Peng

University of Pittsburgh, Pittsburgh, PA, USA
{mingmao,ypeng}@mail.sis.pitt.edu

Abstract. This paper summarizes the results of PRIOR system, which is an ontology mapping system based on Profile pRopagation and InfOrmation Retrieval techniques, for OAEI 2006 campaign. The PRIOR system exploits both linguistic and structural information to map small ontologies, and integrates Indri search engine to process large ontologies. The preliminary results of the experiments for four tasks (i.e. benchmark, web directories, anatomy and food) are presented. A discussion of the results and future work are given at the end.

1 Presentation of the system

1.1 State, purpose, general statement

The World Wide Web (WWW) makes a large number of digital resources publicly accessible. However, finding relevant information, i.e. searching for digital resources from various sources and manually organizing them for relevance, becomes more and more intractable. Semantic interoperability research is aimed at enabling different information systems to communicate information consistently with the intended meaning. Ontology mapping is one critical mechanism to achieve semantic interoperability.

Different communities have proposed different approaches to ontology mapping. The techniques that have been applied to solve mapping problems include linguistic analysis of terms [5][11], comparison of graphs corresponding to the structures [11], mapping to a common reference ontology [4], use of heuristics that look for specific patterns in the concepts definitions [10][8][12][9], and machine-learning techniques [7][2][3][1].

Our approach begins with the belief that the combination of linguistic analysis and graph theory will lead to successful mapping. It explores information from two perspectives, linguistic and structural, to determine the correspondences that identify similar elements in different ontologies. In an ontology, linguistic information is the descriptive information, such as name (i.e. ID), label, comment and property restriction, of a concept (i.e. class, individual and property). Structural information refers to relationships between concepts in the ontology. Such relationships include hierarchy relation, inverse relation and so on. Since the filed of information retrieval

is highly relevant to ontology mapping, we also explore using classic information retrieval method to support the mapping of large ontologies. Figure 1 depicts the architecture of PRIOR system. The details of the approach are explained in next section.

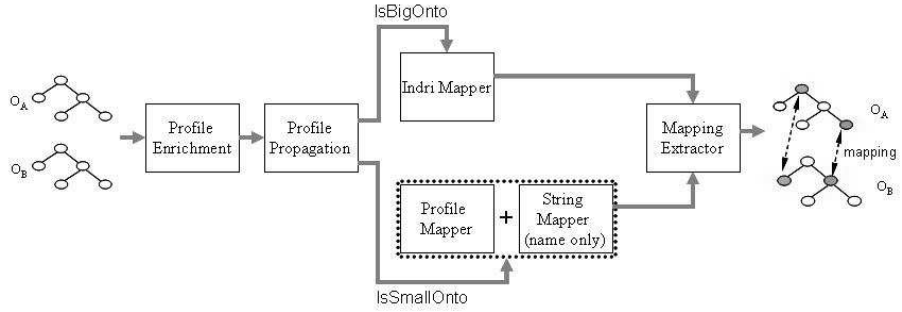

Figure 1 The architecture of PRIOR system

1.2 Specific techniques used

We introduce the term "profile". Similar to the virtual document used in Falcon-AO system [11], the profile of a concept is a combination of all linguistic information of the concept, i.e. the profile of a concept = the concept's name + label + comment + property restriction + other descriptive information. The Profile Enrichment is a process of using a profile to represent a concept in the ontology, and thus enrich its information. The purpose of profile enrichment is based on the observation that though a name is always used to represent a concept, sometimes the information carried in a name is restricted. While, other descriptive information such as comments may contain words that better convey the meaning of the concept.

The Profile Propagation exploits the neighboring information of each concept. That is, we pass the profile of the ancestors, children or siblings of the concept to the profile of the concept itself. The reason why we do profile propagation is based on the observation that if we see the taxonomic tree of an ontology as the index of a book, the super class in the ontology reflects the "context" of its subclasses and each subclass is the "content" of its super class. The process of profile propagation can be represented as: $V_{Nnew} = \sum_{N' \in S} w(N, N')V_{N'}$, where N and N' represent two concepts in the ontologies, S represents the set of all concepts in the ontologies, V_{Nnew} represents the new profile vector of the concept N, $V_{N'}$ represents the profile vector of the concept N', and w(N, N') is a function that assigns different weights to the neighbors of the concept according to the distance between them. Two principles to assign the weight are applied: 1) The closer the two concepts are, the higher weight will be assigned, i.e. the weight of a parent is higher than the weight of a grandparents and the weight of a child is higher than the weight of a grandchild. 2) The weight of a parent is higher than the weight of a child and the weight of a child is higher than the

weight of a sibling. This is because children inherit all characteristics of the parent and may extend some characteristics that parent doe not have, and sibling is usually a complementary of the concept.
For small ontologies, the Profile Mapper compares each concept of the ontologies by computing cosine similarity of the profile of each concept. Simultaneously, the String Mapper computes the similarity between the names of different concepts using Levenshtein distance. The profile similarity and the name string similarity are further integrated to obtain final similarities between concepts. However, if the ontology is too large, calculating the similarity matrix will require too many computing resources and it is time consuming. Based on the understanding that ontology mapping is also an information retrieval task, we turn to classic information retrieval method to solve the problem. Specifically, we integrated indri[1] search engine into PRIOR system. First, the Indri Mapper uses Indri to index profiles of concepts in ontology A. Then queries are generated based on the profiles of the concepts in ontology B. After storing the top-ranked results returned by the queries, we switch two ontologies, i.e. this time ontology B is indexed and queries are generated based on ontology A. The Indri Mapper will pass two sets of search results to the Mapping Extractor.
Having the similarity matrix obtained from small ontologies or Indri search results from large ontologies, the Mapping Extractor extracts all candidates of matched concepts and output the results in desired format.

1.3 Adaptations made for the evaluation

We didn't do any major adaptations in order to align the OAEI campaign ontologies. However, for food test, we treat <skos:broader> and <skos:narrower> as parent and child relations.

1.4 Link to the system and parameters file

The system is available at: http://www.sis.pitt.edu/~mingmao/om06/

1.5 Link to the set of provided alignments (in align format)

The result file can be downloaded from http://www.sis.pitt.edu/~mingmao/om06/result.zip

2 Results

In this section we present the results of alignment experiments on OAEI 2006 campaign. All tests are run on a stand-alone PC running Fedora 4 operating system.

[1] http://www.lemurproject.org/indri

The PC has Pentium 4, 3.0GHz processor, 1G memory, 100GB Serial ATA hard disk and SUN JAVA VM 1.5.0_06.

2.1 benchmark

The benchmark tests can be divided into two types. Test 101-266 are systematically generated from reference ontology, in which some information are discarded, and test 301-304 are real bibliographic ontologies. Since our approach is relied on the linguistic information, we obtain high precision and recall where the test ontologies contain the same names (or name conventions) and/or comments as the reference ontology (i.e. test 101, 103, 104, 203, 204, 208, 221-247). However our approach fails in the recall where both name and comments are replaced or missing in the test ontologies (i.e. test 202, 248-266). For tests 201, 206-207 and 210, though the class name has been "removed" or expressed in another language, we can find some matched classed and properties due to the information of comments and instances. For tests 205 and 209 having name synonyms, the performance of our approach is not good because we do not use thesaurus. For real ontologies 301-304, they cover the same domain as reference ontology using similar descriptive information and different structural information. The result of these real tests shows the average performance of our approach is around 80%. The full result of all tests can be found in Appendix.

2.2 directory

The directory real world case consists of aligning web sites directory. It has 4640 elementary tests. Each of them is represented by pairs of OWL ontologies, where classification relation is modeled as OWL subClassOf. Therefore all OWL ontologies are taxonomies, i.e., they contain only classes (without Object and Data properties) connected with subclass relation. We use the same set of parameters and approach as those of benchmark test to obtain alignment results.

2.3 anatomy

The anatomy task is to find alignment between classes in two medical ontologies, FMA ontology and OpenGALEN ontology. FMA has 72559 classes and OpenGALEN has 9564 classes. Due to the huge size of the ontologies, we use Indri approach. Finally 2583 pairs of candidates have been found within 9 minutes.

2.3 food

The food thesaurus mapping task requires to create alignment between the SKOS version of the United Nations Food and Agriculture Organization (FAO) AGROVOC thesaurus, which has around 16000 terms and is expressed in multilingual, and the United States National Agricultural Library (NAL) Agricultural thesaurus, which has

around 41000 terms and is expressed in monolingual. AGROVOC has 28179 concepts, and NAL has 41594 concepts. Due to the similar reason as anatomy task that the size of food thesaurus is too large, we use Indri approach. Finally 11511 pairs of candidates have been found within 73 minutes. Although “narrowMatch” and “broadMatch” are allowed, we can only get “exactMatch”.

3 General comments

3.1 Comments on the results

Since our approach relies on linguistic information such as name, label, comment, and other descriptive information, it can not handle pure graph matching task, like test 248-266 in benchmarks. Also we do not use external resources like WorldNet to process synonyms, which we believe is important in real cases. Furthermore, some ontology like AGROVOC contains labels in foreign languages; currently we do not use this type of information.
We use Alignment API to parse ontologies and generate alignments. When processing FMA ontology in anatomy test, the API reads each owl:Class as a class first and then as an individual one more time. In all properties of a class, only “ID” and “label” are assigned to the class, all other properties such as “part” and “constitutional_part” are assigned to the individual. Since only classes are alignment candidates, we miss all information in individual.

3.2 Discussions on the way to improve the proposed system

One possible improvement is to integrate external resources to increase recall. For instance, WordNet can be integrated to process synonyms and dictionaries can be used to process foreign languages. Another possible improvement is to find out a better way to adjust the propagation weights. It’s possible to train the weights with some training data.

3.3 Comments on the OAEI 2006 test cases

The ontologies in anatomy and food tests are very large and in a different format (i.e. SKOS, Protégé exported RDF) other than benchmark tests. It will be better to have a small part of ontology as training ontology, for which alignments are provided to participants. So that participants can train their approach on this training ontology. We also would like to see the OAEI 2006 campaign to be the first one to provide reference alignment for real word large scale ontologies so that different approaches can be judged in systematic way.

3.4 Comments on the OAEI 2006 measures

Considering the mapping relations in food track, the evaluation process is more complex. If concept A is an “exactMatch” to concept B, and concept C is a “broader” concept of B, then we can say concept A and C has a “broadMatch” relation. First we don’t know whether A-exactMatch-B and A-broadMatch-C will both appear in reference alignment. Second, if they both appear in reference alignment, but only A-exactMatch-B mapping is in an answer alignment, how do we calculate recall regarding A-broadMatch-C mapping?

4 Conclusion

In this paper, we briefly present a system for ontology mapping – PRIOR system, in which we explore linguistic and structural information and profile propagation method to process small ontologies. We also integrate classic information retrieval method to process large ontologies. The preliminary results are carefully analyzed and some future work are discussed.

References

1. Dhamankar, R., Y. Lee, et al. (2004). "iMAP: Discovering Complex Semantic Matches between Database Schemas." Proceedings of the International Conference on Management of Data (SIGMOD).
2. Doan, A., P. Domingos, et al. (2001). Reconciling Schemas of Disparate Data Sources: A Machine-Learning Approach. SIGMOD Conference.
3. Embley, D. W., D. Jackman, et al. (2001). Multifaceted Exploitation of Metadata for Attribute Match Discovery in Information Integration. Workshop on Information Integration on the Web.
4. Gruninger, M. and J. Kopena (2003). "Semantic integration through invariants." Workshop on Semantic Integration at ISWC.
5. Hovy, E. (1998). Combining and standardizing large-scale, practical ontologies for machine translation and other uses. In The First International Conference on Language Resources and Evaluation (LREC), Granada, Spain.
6. Jaro, M. A. 1989. Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida. Journal of the American Statistical Association 84:414–420.
7. Li, W., C. Clifton, et al. (2000). "Database integration using neural network: implementation and experience." Knowledge and Information Systems 2(1): 73-96.
8. Madhavan, J., P. Bernstein, et al. (2001). Generic Schema Matching Using Cupid. Proceedings of the 27th International Conference on Very Large Data Bases (VLDB), Roma (IT), Morgan Kaufmann Publishers Inc.
9. Melnik, S., H. Garcia-Molina, et al. (2002). Similarity flooding: a versatile graph matching algorithm and its application to schema matching. Proc. 18th International Conference on Data Engineering (ICDE), San Jose (CA US).
10. Mitra, P., G. Wiederhold, et al. (1999). Semi-automatic integration of knowledge sources. Proc. of the 2nd Int. Conf. On Information FUSION'99.

11. Qu, Y., Hu, W., and Cheng, G. 2006. Constructing virtual documents for ontology matching. In Proceedings of the 15th International Conference on World Wide Web (Edinburgh, Scotland, May 23 - 26, 2006). WWW '06. ACM Press, New York, NY, 23-31.
12. Rahm, E. and P. Bernstein (2001). "A survey of approaches to automatic schema matching." The VLDB Journal 10(4): 334-350.

Appendix: Raw results

Matrix of results

#	Name	Prec.	Rec.
101	Reference alignment	1	1
102	Irrelevant ontology	0.00	NaN
103	Language generalization	1	1
104	Language restriction	1	1
201	No names	0.94	0.32
202	No names, no comments	0.6	0.03
203	No comments (was misspelling)	1	1
204	Naming conventions	1	0.94
205	Synonyms	0.63	0.42
206	Translation	0.96	0.7
207		0.96	0.7
208		1	0.93
209		0.53	0.3
210		0.94	0.53
221	No specialisation	1	1
222	Flatenned hierarchy	1	1
223	Expanded hierarchy	1	1
224	No instance	1	1
225	No restrictions	1	1
228	No properties	1	1
230	Flattened classes	0.94	1
231*	Expanded classes	1	1
232		1	1
233		1	1
236		1	1
237		1	1
238		1	1
239		0.97	1
240		0.97	1
241		1	1
246		0.97	1

247		0.97	1
248		0.33	0.01
249		0.6	0.03
250	Individual is empty	1	0.06
251		0.6	0.03
252		0.5	0.03
253		0.33	0.01
254		NaN	0
257		1	0.06
258		0.6	0.03
259		0.5	0.03
260		0.5	0.03
261		0.5	0.03
262		NaN	0
265		0.5	0.03
266		0.5	0.03
301	Real: BibTeX/MIT	0.92	0.74
302	Real: BibTeX/UMBC	0.86	0.63
303	Real: Karlsruhe	0.68	0.82
304	Real: INRIA	0.95	0.96