The PRIOR+: Results for OAEI Campaign 2007

Ming Mao, Yefei Peng

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

Abstract. Ontology mapping is to find semantic correspondences between similar elements of different ontologies. It is critical to achieve semantic interoperability in the WWW. This paper summarizes the results of the PRIOR+ participating at OAEI campaign 2007. The PRIOR+ is a generic and automatic ontology mapping tool, based on propagation theory, information retrieval technique and artificial intelligence model. The approach utilizes both linguistic and structural information of ontologies, and measures the profile similarity of different elements of ontologies in a vector space model (VSM). Furthermore, the PRIOR+ adaptively aggregate different similarities according to the harmony of similarity matrix. Finally the PRIOR+ deals with ontology constraints using interactive activation and competitive neural network. The preliminary results of benchmark task are presented, followed by a discussion. Some future works are given at the end.

1 Presentation of the system

1.1 State, purpose, general statement

The World Wide Web (WWW) now is widely used as a universal medium for information exchange. Semantic interoperability among different information systems in the WWW is limited due to information heterogeneity, and the non semantic nature of HTML and URLs. Ontologies have been suggested as a way to solve the problem of information heterogeneity by providing formal and explicit definitions of data. They may also allow for reasoning over related concepts. Given that no universal ontology exists for the WWW, work has focused on finding semantic correspondences between similar elements of different ontologies, i.e., *ontology mapping*. Automatic ontology mapping is important to various practical applications such as the emerging Semantic Web [3], information transformation and data integration [2], query processing across disparate sources [7], and many others [4].

Ontology mapping can be done either by hand or using automated tools. Manual mapping becomes impractical as the size and complexity of ontologies increases. Fully or semi-automated mapping approaches have been examined by several research studies, e.g., analyzing linguistic information of elements in ontologies [15], treating ontologies as structural graphs [12], applying heuristic rules to look for

specific mapping patterns [8] and machine learning techniques [1]. More comprehensive surveys of ontology mapping approaches can be found in [9][14].

This paper proposes a new generic and scalable ontology mapping approach, the PRIOR+ approach. The architecture of the PRIOR+ is shown in **Fig. 1**. The PRIOR+ takes advantage of propagation theory, information retrieval technique and artificial intelligence model to solve ontology mapping problem. It utilizes both linguistic and structural information of ontologies, and measures the profile similarity of different elements of ontologies in a vector space model (VSM). Finally, the PRIOR+ adaptively aggregates different similarities according to the harmony of the matrix and deals with ontology constraints using interactive activation network.

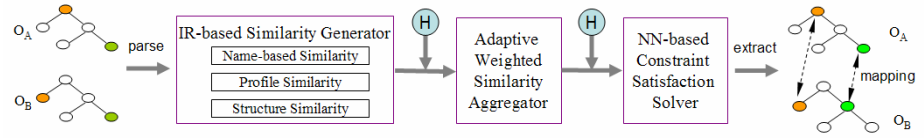

Fig. 1. The architecture of the PRIOR+ approach

1.2 Specific techniques used

The PRIOR+ is extended from the PRIOR [10][11]. In addition to the profile similarity and the edit distance of elements' name used in the PRIOR, the PRIOR+ considers structure similarity as well and adaptively aggregate different similarities based on their harmony. Furthermore, the PRIOR+ has a brand new NN-based Constraint Satisfaction Solver.

1.2.1 Similarity Generation

The similarity generation model aims to generate the similarity of both linguistic and structural information of ontologies. The details of calculating profile similarity and the edit distance of elements' name have been presented in the PRIOR [10][11]. To calculate the structure similarity of two elements, various structural features are extracted, e.g. the number of its sub-elements, the number of its direct property, the depth of the element to the root etc. Afterwards, the difference between these structural features are calculated and normalized to represent its structure similarity. The outputs of the similarity generation model are three similarity matrixes. Each matrix denotes a kind of similarity of two ontologies.

1.2.2 Harmony Estimation

The heterogeneities of information result in differences between ontologies, either from a linguistic view or structural view. Therefore, given two ontologies, it is critical to estimate the difference between ontologies, and then to adjust mapping strategies

according to the difference. Here we define a term called *harmony* to represent the similarity between ontologies. Three types harmony of ontologies, i.e. name harmony, profile harmony and structure harmony, are calculated based on the similarity matrixes output from similarity generation model.

Ideally, if two ontologies are very similar in either linguistic or structural view, two true should-be-mapped elements should own a similarity equal to 1 or larger than the similarity of all other cells standing in the same row and column of those two elements in the corresponding similarity matrix. Therefore, the harmony of ontologies can be defined using Equation 1, where h_k denotes different types of harmony (i.e., name harmony, profile harmony and structure harmony), E_{O_1} and E_{O_2} denote the number of elements in ontologies, O_1 and O_2, $CMAX_{M_k}$ denotes the number of cells that own the highest similarity in its corresponding row/column in similarity matrix M_k.

$$h_k = \frac{\#CMAX_{M_k}}{\min(\#E_{o_1}, \#E_{o_2})} \quad (1)$$

The different harmony of ontologies are used as weights to adaptively aggregate name similarity, profile similarity and structure similarity output from similarity generation model. Finally, the harmony of the aggregated similarity is estimated using the same way. The final harmony, h_f, will decide the necessity of NN-based Constraint Satisfaction Solver. If $h_f > c$ (c is an experience number), the cells having largest similarity in each row/column will be output to NN-based Constraint Satisfaction Solver as refined hypotheses. Otherwise, all cells in the final similarity matrix will be output.

1.2.3 NN-Based Constraint Satisfaction Solver

Constraint satisfaction problem (CSP) [16] arises as an intriguing research problem in ontology mapping due to the characteristics of ontology itself and its representations. The hierarchical relations in RDFS, the axioms in OWL and the rules in SWRL result in different kinds of constraints. For example, "if concept A matches concept B, then the ancestor of A can not match the child of B in the taxonomy" and "two classes match if they have owl:sameAs or owl:equvalentClass relations". To improve the quality of ontology mapping, it is critical to find the best configuration that can satisfy such constraints as much as possible.

CSPs are typically solved by a form of search, e.g. backtracking, constraint propagation, and local search [16]. The interactive activation network is first proposed to solve CSPs in [13]. The network usually consists of a number of competitive nodes connected to each other. Each node represents a hypothesis. The connection between two nodes represents constraint between their hypotheses. Each connection is associated with a weight. For example, we have two hypotheses, H_A and H_B. If whenever H_A is true, H_B is usually true, then there is a positive connection from node A to node B. Oppositely if H_A provides evidence against H_B, then there is a negative connection from node A to node B. The importance of the constraint is proportional to the strength (i.e. *weight*) of the connection representing that constraint. The state of a

node is determined locally by the nodes adjacent to it and the weights connecting to it. The state of the network is the collection of states of all nodes. Entirely local computation can lead the network to converge to a global optimal state.

In the context of ontology mapping, a node in an interactive activation network represents a hypothesis that element E_{1i} in ontology O_1 can be mapped to element E_{2j} in ontology O_2. The initial activation of the node is the similarity of (E_{1i}, E_{2j}) output from the adaptive similarity aggregation model. The activation of the node can be updated using the following simple rule, where a_i denotes the activation of *node i*, written as n_i, net_i denotes the net input of the node.

$$a_i(t+1) = \begin{cases} a_i(t) + net_i(1 - a_i(t)), & net_i > 0 \\ a_i(t) + net_i a_i(t), & net_i < 0 \end{cases} \quad (2)$$

The net_i comes from three sources, i.e. its neighbors, its bias, and its external inputs, as defined in Equation 3, where w_{ij} denotes the connection weight between n_i and n_j, a_j denotes the activation of node n_j, $bias_i$ denotes the bias of n_i, ei_i denotes the external input of n_i, which is a function of the confidence of a mapping, *istr* and *estr* are constants that allow the relative contributions of the input from internal sources and external sources to be readily manipulated. Note that the connection matrix is symmetric and the nodes may not connect to themselves, i.e., $w_{ij}=w_{ji}$, $w_{ii}=0$.

$$net_i = istr \times (\sum_j w_{ij} a_j + bias_i) + estr \times (ei_i) \quad (3)$$

Furthermore, the connections between nodes in the network represent constraints between hypotheses. For example, the constraint that "only 1-to-1 mapping is allowed" results in a negative connection between nodes (E_{1i}, E_{2j}) and (E_{1i}, E_{2k}), where $k \neq j$. Moreover, "two elements match if their children match", results in a positive connection between nodes (E_{1i}, E_{2j}) and (E_{1k}, E_{2t}), where E_{1k} and E_{2t} are the children of E_{1i} and E_{2j} respectively. Finally, the complexity of the connections may be very large because of complex constraints.

1.3 Adaptations made for the evaluation

We didn't make any specific adaptations for the tests in the OAEI campaign 2007. All the mappings output by the PRIOR+ are based on the same set of parameters.

1.4 Link to the system and parameters file

The PRIOR+ is available at: http://www.sis.pitt.edu/~mingmao/om07/.

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/om07/priorplus.zip

2 Results

In this section we present the results of the PRIOR+ in OAEI campaign 2007. All tests are run on a stand-alone PC running Ubuntu 6.0.6 operating system. The PC has Intel Dual Core 1.8 Hz processor, 1.5G memory, 100GB Serial ATA hard disk and SUN JAVA VM 1.6.0.

2.1 Benchmark

The benchmark track is the only track that opens its ground truth for participants. According to the different characteristics of ontologies, most parameters of the PRIOR+ are tuned on it. The full result of all tests can be found in Appendix[1].

The results show that Test 101, 103 and 104 are perfect because all names, comments and instances of classes and properties are the same. Test 201-210 are very structurally similar as the reference ontology, therefore the structural harmony plays an important role in deciding the final similarity of the elements of ontologies. Test 221-247 have high linguistic similarity with reference ontology, and thus the PRIOR+ obtained good performance on it. Test 248-266 are both linguistic and structural different with reference ontology. Even with the usage of the structural information, the PRIOR+ has some improvement compared with the PRIOR. The recall of these tests is still a little bit low. The reason why the PRIOR+ did not work well in these tests is under investigation. The 301-304 are real world ontologies, which have more impact when evaluating the mapping approach. The PRIOR+ also gained good results in all these tests.

Meanwhile, test 202, 209, 210, 248-266 and real case 302 and 303 demonstrate the effectiveness of using the interactive activation network to solve constraint satisfaction problem in ontology mapping.

2.2 Other Tracks

The web directory, anatomy and food track are all blind tracks that means no ground truth is available for participants to analyze the performance of the proposed approach. Therefore, please refer to the final results published by OAEI for further information.

[1] The data presented is slightly different from what we submitted to the OAEI campaign 2007 after improving the PRIOR+ approach.

3 General comments

3.1 Discussions on the way to improve the proposed system

Parameter tuning is an important issue in the implementation of neural network in our future work. Another possible improvement is to integrate auxiliary information and Web information for ontology mapping. For example, auxiliary information such as WordNet can be used to process synonyms. The co-occurrence of two elements returned by search engines can contribute to identify their semantic relation.

3.2 Comments on the OAEI 2006 test cases

Currently most tests in the campaign are blind. It will be better for OAEI to provide a small part of ground truth in some tests, such as anatomy, for participants to explore machine learning techniques. Meanwhile, in web directory task, some loops existing in the test cases have been broken randomly in the implementation of the PRIOR+.

4 Conclusion

In this paper, we present the PRIOR+, a generic ontology mapping tool, and its results in OAEI campaign 2007. The PRIOR+ integrates propagation theory, information retrieval technique and the interactive activation network to solve ontology mapping problem. The preliminary result of the PRIOR+ in benchmarks tests is promising.

References

1. Doan, A., J. Madhaven, et al. (2003). "Learning to Match Ontologies on the Semantic Web." VLDB Journal **12**(4): 303-319.
2. Dou, D., D. McDermott, et al. (2005). "Ontology Translation on the Semantic Web." Journal on Data Semantics (JoDS) **II**: 35-57.
3. Ehrig, M. (2006). Ontology Alignment: Bridging the Semantic Gap (Semantic Web and Beyond). ISBN-038732805X. Springer. 2006.
4. Euzenat, J., Bach, T., et al. (2004). State of the art on ontology alignment, Knowledge web NoE.
5. Euzenat, J et al. (2006). Results of the Ontology Alignment Evaluation Initiative 2006. In Proceedings of ISWC 2006 Ontology Matching Workshop. Atlanta, GA.
6. Felzenszwalb, P. F. and Huttenlocher, D. P. (2006). Efficient belief propagation for early vision. International Journal of Computer Vision, Vol. 70, No. 1.
7. Gasevic, D. and M. Hatala (2005). "Ontology mappings to improve learning resource search." British Journal of Educational Technology.

8. Hovy, E. (1998). Combining and standardizing large-scale, practical ontologies for machine translation and other uses. In Proceedings of the 1st International Conference on Language Resources and Evaluation (LREC), Granada, Spain.
9. Kalfoglou, Y. and M. Schorlemmer (2003). "Ontology mapping: the state of the art." The Knowledge Engineering Review 18(1): 1-31.
10. Mao, M. and Peng, Y. (2006). PRIOR System: Results for OAEI 2006. In Proceedings of ISWC 2006 Ontology Matching Workshop. Atlanta, GA.
11. Mao, M., Peng, Y. and Spring, M. (2007) A Profile Propagation and Information Retrieval Based Ontology Mapping Approach, In Proceedings of SKG 2007.
12. 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).
13. McClelland, J. L. and Rumelhart, D. E. (1988). Explorations in Parallel Distributed Processing: A Handbook of Models, Programs, and Exercises. The MIT Press.
14. Noy, N. (2004). "Semantic Integration: A Survey of Ontology-Based Approaches." SIGMOD Record **33**(4): 65-70.
15. 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.
16. Tsang, E. (1993). Foundations of Constraint Satisfaction: Academic Press.

Appendix: Raw results

Matrix of results

algorithm		prior+	
Test #	Precision	Recall	F-Measure
101	1	1	1
103	1	1	1
104	1	1	1
201	1	1	1
202	0.9756	0.82	0.894
203	1	1	1
204	1	1	1
205	0.9688	0.96	0.964
206	1	0.99	0.995
207	1	0.99	0.995
208	1	0.96	0.979
209	0.8919	0.68	0.772
210	0.9634	0.81	0.883
221	1	0.98	0.99
222	1	0.96	0.978
223	1	1	1
224	1	1	1

225	1	1	1
228	1	1	1
230	0.9351	1	0.966
231	1	1	1
232	1	1	1
233	1	1	1
236	1	1	1
237	1	1	1
238	1	1	1
239	0.9667	1	0.983
240	0.9706	1	0.985
241	1	1	1
246	0.9667	1	0.983
247	0.9706	1	0.985
248	0.9143	0.66	0.767
249	1	0.84	0.91
250	0.8065	0.76	0.781
251	0.9531	0.66	0.777
252	0.8904	0.67	0.765
253	0.913	0.65	0.759
254	1	0.27	0.429
257	0.6774	0.64	0.656
258	0.9219	0.63	0.752
259	0.8904	0.67	0.765
260	0.7895	0.52	0.625
261	0.4333	0.39	0.413
262	1	0.27	0.429
265	0.7368	0.48	0.583
266	0.5	0.45	0.476
301	0.9259	0.82	0.87
302	0.9677	0.63	0.76
303	0.82	0.84	0.828
304	0.9136	0.97	0.943
H-mean	0.9577	0.87	0.912