Cluster-based Similarity Aggregation for Ontology Matching

Quang-Vinh Tran[1], Ryutaro Ichise[2], and Bao-Quoc Ho[1]

[1] Faculty of Information Technology, Ho Chi Minh University of Science, Vietnam
{tqvinh,hbquoc}@fit.hcmus.edu.vn
[2] Principles of Informatics Research Division, National Institute of Informatics, Tokyo, Japan
ichise@nii.ac.jp

Abstract. CSA is an automatic similarity aggregating system for ontology matching. The systems have two main part. The first part is calculation and combination of different similarity measures. The second part is the alignment extraction. The system first calculate five different basic measures to create five similarity matrix, i.e, String based similarity measure, WordNet based similarity measure... Next, it exploits the advantage of each measure through a weight estimation process. Then these similarity matrix is combined into a final similarity matrix. After that, the pre-alignment is extract from this matrix. Finally the pruning process is applied to increase the accuracy of the system.

1 Presentation of the system

Ontologies are widely use to provide semantic to the data in the new internet environment. Since they are created by different user for different purpose, It is necessary to develop a method to match multiple ontologies for integrating data from different resources [2].

1.1 State, purpose, general statement

CSA (**C**luster-based **S**imilarity **A**ggregation) is the automatic weight aggregating system for ontology alignment. The system is designed to search for semantic correspondence between heterogeneous data sources from different ontologies. The current implementation only support one-to-one alignment between concepts and properties (including object properties and data properties). The core of CSA is the utilizing the advantage of each basic strategy for the alignment process. For example, the String based similarity measure works well in case the two entities are similar in linguistic while Structure based similarity measure is effective when the two entities have similar in their local structure. The system automatically combines many similarity measurements based on the analysis of their similarity matrix. Detail of the system is described in the following part.

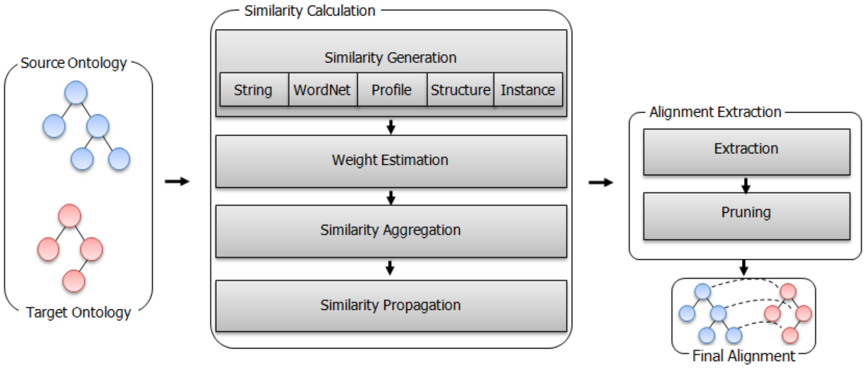

Fig. 1. The main process of CSA

1.2 Specific techniques used

The process of the system is illustrated in Figure 1. First, we calculate five basic similarity measures. These similarities are String edit distance, WordNet based, Profile, Structure, and Instance based. Second, the weight for each similarity is estimated through a weight estimation process. Then, we aggregated these similarities based on their weights. After that, we propagate the similarity to get the final similarity matrix. The pre-alignment is then extracted from this matrix. Finally, we apply the pruning process to get the final alignment.

Similarity Generation The similarity between entities in the two ontology is computed by five basic measures. The String edit distance measures the lexical feature of the entity's name. The WordNet [3] exploits the similarity between words occur in the entity's name. We use the method of Wu and Palmer for calculating the WordNet similarity [8]. The Profile similarity makes use of the id, label, comments information contained in entity. The profile for a class take their properties, instances into account. The profile for a property includes their domains and their ranges. Then we construct the weight feature vector using tf-idf. The similarity is then calculated by the cosine similarity of the two vectors. The Structure similarity is calculated for class only. This similarity measures the difference in the local structure of an entity. We implement the method introduced in [7] for the structure measure. This calculation is based on the difference of number of class's children, number of class's sibling, the normalized depth from root and the number of properties restricted to this class. The Instance based measure is similar to the Profile except that we only utilize the content of instances belong to classes and the properties appear in these instances.

Weight estimation Weight estimation is the core of CSA. In this step, we analyse each similarity matrix of each basic measure to find which one is actually effective and which one is not for alignment process. The weight estimation process is based on two information. First, for each single method, the process of finding a threshold for distinguish matching pairs from non matching pairs can be viewed as a binary classification problem [5]. In this process the positive class contains matching pairs and the negative class contains non matching ones. Second, in one-to-one ontology alignment, the maximum number of matching pairs is equal to the minimum number of entities in the two ontologies. Then if a single method is effective, its correspondent similarity matrix must have the two criteria: The matrix must have the ability of distinguishing matching pairs from non matching pairs and the number of matching pairs must approximate the minimum number of entities in the two ontologies.

Base on these criteria, we model the weight estimation process for concept as follow: First, for each similarity matrix we use K-means algorithm to cluster the similarity values into two different classes (k = 2). The feature is the similarity value of each pairs of classes. The cluster with higher mean represents the matching set, and the lower one represents the non matching set. Then we filter out all the values belong to the non matching set. What remains is the similarity matrix with the higher values. Then we calculate the number of row that has value in the matrix. These row represent the possible matching pairs. Because in our case we consider the one-to-one matching, then one concepts from source ontology is only matched up to one concepts from target ontology. Finally, the weight is estimated by the ratio of the number of row over the number of matched values in the filtered matrix.

$$weight = \frac{|\textit{number of row that has value}|}{|\textit{number of value in matching set}|} \tag{1}$$

The weight estimation for property similarity matrix is calculated in the same manner.

Similarity Aggregation The similarity combination can be defined as the weight average of the five basic measures. The weight for each measure is estimated in the previous step.

$$Sim_{combine}\left(e_1, e_2\right) = \frac{\sum_{i=1}^{n} weight_i \times Sim_i\left(e_1, e_2\right)}{\sum_{i=1}^{n} weight_i} \tag{2}$$

Similarity Propagation This step consider the impact of structural information on the similarity between entity pair in the aggregated matrix. The intuition is that the more similar in structure two entities are, the more similar they are. To exploit the structure information, we use the method of Descendant Similarity Inheritance [1].

Extraction In our system, only one-to-one matching is allowed. The final similarity matrix can be viewed as a bipartite graph with the first set of vertices are entities from source ontology and the second set of vertices are entities from target ontology. Thus, the alignment extraction can be modelled as the process of finding the mapping from the bipartite graph. To solve this, we apply the stable marriage problem algorithm [4]. We models the two set of entities as sets of men and women. For each man and each woman, in the correspondence set, a list of priority of men and women is created based on their similarity value. The stable marriage algorithm is then applied to find the stable mapping between two sets. The result is the pre-alignment.

Table 1. The performance of CSA on the benchmark track

Ontology	Prec.	Rec.
101	1.0	1.0
201 - 202	0.84	0.71
221 - 247	0.97	0.99
248 - 252	0.76	0.63
253 - 259	0.77	0.57
260 - 266	0.62	0.51
Average	0.79	0.66

Pruning This is the final step of our system. In this step we filter out a proportion of entities pair that have low confidence to increase the precision of our system. For the threshold, we set it manually. The result is the final alignment of the two ontologies.

1.3 Adaptations made for the evaluation

We do not make any specific adaptation for the OAEI 2011 campaign. The three track are run in the same set of parameter.

1.4 Link to the system and parameters file

The CSA system can be downloaded from seal-project at `http://www.seals-project.eu/`.

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

The result of CSA system can be downloaded from seal-project at `http://www.seals-project.eu/`.

2 Results

In this section, we present the results of the CSA system. We participate in tree track of benchmarks, anatomy and conference. The result is in the following part.

2.1 Benchmarks

On the benchmarks of 2011, the reference ontology are different from the previous year. Since the descriptions, restrictions and instances are limited, it affects our algorithm very much. The result is shown at Table 1. We group the test into six groups based on their difficulty. The result shows the average precision and recall for each group.

Table 2. The performance of CSA on the anatomy track

Prec.	Rec.

Table 3. The performance of CSA on the conference track

Test	Prec.	Rec.

2.2 Anatomy

The anatomy dataset consist of two large ontology of Adult Mouse Anatomy with 2247 classes and a part of NCI Thesaurus for describing human anatomy with 3304 classes. CSA result is shown at Table 2. Because of the high cost of computation, the execution time is quite high.

2.3 Conference

The results of conference are illustrated in Table 3. It is difficult to archive the good results, because ontologies from this track are real and developed by different organizations for different purposes.

3 General comments

This is the first time CSA participating in the OAEI tracks, and our systems is new to the seals platform. Further, the same set of parameter for all test in all tracks are difficult, because for each tracks the ontologies has different characteristic to be processed. Then, in different dataset we need different method for defining the threshold to extract the final alignment.

3.1 Comments on the results

Strengths CSA can automatically combine different similarity measures. Our system do not need any external resources or training data for estimate the weight in the aggregation step.

Weaknesses The structure based similarity included in CSA is not strong enough to distinguish the different between matching pairs and non matching pairs. There are no structure similarity for properties. Further, we do not integrated any semantic verification or constraint in our system yet.

3.2 Discussions on the way to improve the proposed system

Our system is new and we still have chance to improve our method. First, we can integrate more basic similarity measures in our system for aggregating. Second, for the pruning step, we can find the way for automatic defining a threshold rather than manually tuning. Finally, we can use some semantic verification as in [6] to pruning the low confidence matching pairs.

4 Conclusion

This is the first time CSA participate in OAEI campaign. In this year, we participate in three tracks of benchmarks, anatomy and conference. We have introduced the new method for aggregating different similarity measures. The results show that our method is promising.

References

1. Isabel F. Cruz and William Sunna. Structural alignment methods with applications to geospatial ontologies. *Transactions in GIS*, 12(6):683–711, 2008.
2. Jérôme Euzenat and Pavel Shvaiko. *Ontology matching*. Springer-Verlag, Heidelberg (DE), 2007.
3. Christiane Fellbaum, editor. *WordNet: an electronic lexical database*. MIT Press, 1998.
4. David Gale and Lloyd S Shapley. College admissions and the stability of marriage. *American Mathematical Monthly*, 69(1):9–15, 1962.
5. Ryutaro Ichise. Machine learning approach for ontology mapping using multiple concept similarity measures. In *Proceedings of the Seventh IEEE/ACIS International Conference on Computer and Information Science*, pages 340–346, Washington, DC, USA, 2008. IEEE Computer Society.
6. Yves R. Jean-Mary, E. Patrick Shironoshita, and Mansur R. Kabuka. Ontology matching with semantic verification. *Web Semantics: Science, Services and Agents on the World Wide Web*, 7:235–251, September 2009.
7. Ming Mao, Yefei Peng, and Michael Spring. An adaptive ontology mapping approach with neural network based constraint satisfaction. *Web Semantics: Science, Services and Agents on the World Wide Web*, 8:14–25, March 2010.
8. Zhibiao Wu and Martha Palmer. Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, ACL '94, pages 133–138, Stroudsburg, PA, USA, 1994. Association for Computational Linguistics.