Towards a Cluster-based Approach for User Participation in Ontology Maching

Vinicius Lopes, Fernanda Baião, Kate Revoredo

Department of Applied Informatics
Federal University of the State of Rio de Janeiro (UNIRIO), Rio de Janeiro, Brazil
{vinicius.lopes,fernanda.baiao,katerevoredo}@uniriotec.br

Abstract. User participation is a promising approach for Ontology Matching; however, determining the most representative pairs of entities is still a challenge. This paper delineates an Ontology Matching approach for user participation employing a clustering algorithm.

Keywords. ontology matching, machine learning, clustering

1 Introduction

Ontology matching focuses on identifying correspondences between entities of two or more ontologies and establishing an alignment as a solution to the heterogeneity problem. Some works in ontology matching apply user participation approaches [2][5], such as selecting and combining similarity measures, tuning parameter values or giving feedback for suggested correspondences. User feedback is considered a promising approach since it requires domain knowledge as opposed to technical knowledge. Due to the difficulty of finding available users, however, it is necessary to minimize user effort by selecting the most representative correspondences. This work delineates an approach to address this issue, in which we apply a clustering algorithm to identify the most representative pairs of entities.

2 A Clustering-based Approach for User Participation

Our proposed approach is composed by 4 steps, which are detailed below.

Select Candidate Correspondences. In this step, a committee is formed to select a subset of candidate correspondences for the user feedback. Given two ontologies O and O´, each committee member m_i is represented by a matrix M_i. Each cell $M_i[x,y]$ is the similarity value (calculated according to a unique or a combination of similarity measures) for the pair (x,y), where x is an entity of O and y is an entity of O´. Since Mi are typically sparse matrices (given that most of the pairs do not match), this step analyzes all matrices and selects pairs with the highest potential for actually being correspondent. A pair (x,y) is selected as a candidate correspondence iff, for every matrix Mi, y is the entity that is most similar to x, and vice-versa.

Select Correspondences for User Feedback. In this step, we apply the algorithm farthest-first [1] as a naïve, yet effective and efficient clustering algorithm for selecting correspondences for user feedback among the candidate correspondences. Each instance to be clustered represents a candidate

correspondence (x, y). The attributes of an instance (x, y) are the similarity values Mi[x][y] of each matrix. The cluster centroids are selected for user feedback and then stored in a repository.

Collect and Propagate User Feedback. The user gives his feedback on the selected pairs (either confirming or rejecting as a real correspondence). The feedbacks are updated in the repository.

Learn the Ontology Alignment and Propagate User Feedback. In this step, a classification algorithm is executed considering the repository of classified correspondences. The Naive Bayes classification algorithm achieved the best results. The bayes rule determines the probability distribution of class *C* for a pair of entities, considering its attributes (similarity measures). The resulting model is used to classify candidate correspondences, returning the label *c* that maximizes the posterior probability to propagating the effect of user feedback for the remaining candidate correspondences, and storing them in the repository.

We executed an initial experiment of the approach on top of the OAEI conference dataset. Reference alignments were used to validate the results and simulate user feedbacks. We considered only equivalence correspondences between classes. The committee included Cosine [4] and WuPalmer [3] similarity measures. We evaluated two values (3 and 6) for the number of clusters, or user feedbacks. In the first run the approach achieved an average precision of 0.68 and an average recall of 0.55. In the second run the approach achieved an average precision of 0.83 and an average recall of 0.58. These results show an increase in the precision of 15% when the number of feedbacks increases. F-measure also increased from 0.58 from 0.67. However, the metrics remained the same (or even decreased) for certain pairs of ontologies, indicating there is a need to further investigate the optimal number of clusters for each case.

3 Conclusion

We introduce an approach for ontology matching with user participation that selects candidate correspondences based on a committee of similarity measures. Promising results were obtained on top of the OAEI conference dataset. Future work will perform further experiments, consider other similarity measures and clustering algorithms (including hierarchical approaches).

References

1. Gonzalez, T. F. Clustering to minimize the maximum intercluster distance. Theoretical Computer Science 38, pp. 293–306 (1985).
2. Cruz, I.F., Stroe, C., Palmonari, M.: Interactive User Feedback in Ontology Matching Using Signature Vectors. In: Kementsietsidis et al (eds.) ICDE. pp. 1321–1324 (2012).
3. Wu, Z., Palmer, M.: Verb Semantics and Lexical Selection. Proc. 32nd annual meeting on Association for Computational Linguistics. pp. 133–138 (1994).
4. Stoilos, G., Stamou, G., Kollias, S.: A String Metric for Ontology Alignment. Semant. Web–ISWC 2005. 3729, 624–637 (2005).
5. Shi, F., Li, J., Tang, J., Xie, G.T., Li, H.: Actively Learning Ontology Matching via User Interaction. In: Bernstein, A. et al (eds.). ISWC. pp. 585–600. Springer (2009).