MappingAssistant: Interactive Conflict-Resolution for Data Integration

Faraz Fallahi[1], Jan Noessner[2]
Eva Maria Kiss[1], and Heiner Stuckenschmidt[2]

[1] ontoprise GmbH,
An der Raumfabrik 29, 76227 Karlsruhe, Germany
{fallahi,kiss}@ontoprise.de
[2] KR & KM Research Group
University of Mannheim, B6 26, 68159 Mannheim, Germany
{jan,heiner}@informatik.uni-mannheim.de

Abstract. Enterprise applications often face the problem of integrating heterogeneous data due to the growing number of distributed systems. Leveraging semantic web technologies with ontologies as target schema for matching data is a succesful approach for data integration. Hence, we developed a new interactive approach in the scope of the MappingAssistant project for identifying errors within alignments. It is based on a diagnostic method combined with human-understandable explanations on the instance level. This supports the users in finding wrong rules or facts in a time-saving manner.

Introduction. Alignments produced by automated ontology matching algorithms are error-prone. Consequently, their results need to be supervised by a human domain expert. As they are usually represented in ways only technical experts can deal with, the evaluation task is complicated and time-consuming. The MappingAssistant simplifies the alignment evaluation process by decreasing the amount of user interactions needed to correct the automatically generated mapping results. In particular, we developed a user-friendly conflict resolution method utilizing data reduction techniques for illustrating representative individuals and an algorithm for minimizing user interaction in the evaluation process of aligned data.

User Interface. In data integration scenarios users are often faced with ill-labeled concepts as for instance cryptographic database labels. As a result, the expert is used to investigate the data on the instance level. Existing applications like the AgreementMaker [1] illustrate alignments mostly on the schema level. Our framework also represents alignments on the concept level through a graph-based view and a table-based view (Figure 1 upper left and upper right). Additionally, we visualize instances (Figure 1 lower right) and natural language based presentations of proof-tree nodes (Figure 1 lower left). The user is able to specify the correctness of the matching either on the concept level, the instance level, or through the help of our diagnostic-based proof-tree dialog. The approach is completely implemented as an extension of the OntoStudio [3] environment.
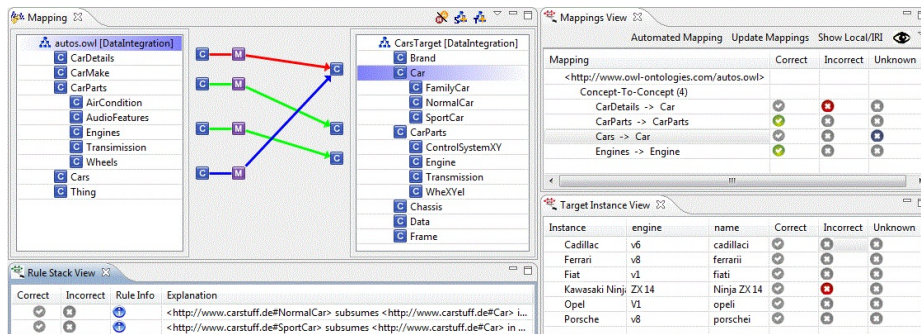
Fig. 1. MappingAssistant Surface

Data Simplification. In order to avoid overwhelming the user by presenting him thousands of instances we combine different clustering algorithms for data simplification. We utilize attribute-driven variations of weighted hierarchical and partial clustering algorithms partially described in [2]. One key step of our approach determines critical outliers which are the most likely to reveal wrong alignments. Clustering algorithms are often specialized either for numerical, categorical. or string-based feature types. Therefore, we split the investigated instance data according to these types into subsets which are then processed by the corresponding cluster algorithm. As a result, one composed representative of each generated cluster is presented to the user.

Proof-Tree Dialog. We developed an approach to minimize the user questions necessary for determining wrong rules or facts. This approach works on a prolog-based proof-tree [4] that depicts an explanation of the result for a given query. In our case the query represents assertions, like *Mammal(Bee)*, which the user selected as a wrong mapping result. Since we assume the correctness of user decisions, the proof-tree of this wrong mapping result must contain at least one wrong node. For finding this wrong node we ask a minimal amount of questions in natural language such as *"Is a bee an animal"*. In particular, we determine the node for the next question in a way that remaining questions are minimized for both, a correct or an incorrect answer of the user.

References

1. I. Cruz, F. Antonelli, and C. Stroe. AgreementMaker: efficient matching for large real-world schemas and ontologies. *Proceedings of the VLDB Endowment*, 2(2):1586–1589, 2009.
2. J. Hair, W. Black, B. Babin, R. Anderson, and R. Tatham. *Multivariate data analysis*, volume 7. Prentice hall Upper Saddle River, NJ, 2009.
3. A. Maier, H. Schnurr, and Y. Sure. Ontology-based information integration in the automotive industry. *The SemanticWeb-ISWC 2003*, pages 897–912, 2003.
4. A. Walker. Prolog/Exl, an inference engine which explains both yes and no answers. In *Proceedings of the Eighth international joint conference on Artificial intelligence-Volume 1*, pages 526–528. Morgan Kaufmann Publishers Inc., 1983.