An integrated matching system: GeRoMeSuite and SMB – Results for OAEI 2010

Christoph Quix[1], Avigdor Gal[2], Tomer Sagi[2], David Kensche[1]

[1]RWTH Aachen University, Germany
http://www.dbis.rwth-aachen.de

[2]Technion – Israel Institute of Technology
http://www.technion.ac.il/

Abstract. We present the results of an integrated matching system which is the result of a cooperation project between the Israel Institute of Technology (Technion) and the RWTH Aachen University in Germany. We have integrated the GeRoMeSuite system (from RWTH Aachen) and SMB (from Technion). Both tools aim at matching schemas; while GeRoMeSuite offers a variety of matchers, SMB provides the information on how to combine matchers and how to enhance match results. Thus, an integration of the tools is beneficial for both systems.

1 Presentation of the system

1.1 GeRoMeSuite

As a framework for model management, *GeRoMeSuite* [3] provides an environment to simplify the implementation of model management operators. *GeRoMeSuite* is based on the generic role based metamodel *GeRoMe* [2], which represents models from different modeling languages (such as XML Schema, OWL, SQL) in a generic way. Thereby, the management of models in a polymorphic fashion is enabled, i.e., the same operator implementations are used regardless of the original modeling language of the schemas. In addition to providing a framework for model management, *GeRoMeSuite* implements several fundamental operators such as Match [6], Merge [5], and Compose [4].

The matching component of *GeRoMeSuite* has been described in more detail in [6], where we present and discuss in particular the results for heterogeneous matching tasks (e.g., matching XML Schema and OWL ontologies). An overview of the complete *GeRoMeSuite* system is given in [3].

1.2 SMB

The Schema Matching Boosting (SMB) Service is a toolkit for enhancing the performance of schema matchers. SMB operates in 3 modes: Enhance, Learn, and Recommend. In the *enhance* mode, SMB recieves a raw correspondence matrix (with similarity values for attribute correspondence in the range of [0,1]) and performs an analysis of the results per row and column. Subsequently, SMB uses contrasting and weakening algorithms to boost results of “promising” rows and columns and weaken results of “non-promising” rows and columns respectively. Contrasting is perfromed using a modified version of the Weber contrast function. Weakening is inversly proportional to the row and column average.

The *learn* mode is used to perform off-line training of SMB on the perfromance behavior of matchers w.r.t. various matching tasks which are classified to classes according to their a-priory features such as schema size. Training is performed using the SMB algorithm, as introduced in [1]. The *recommend* classifies in run-time a given matching task, providing the reccomended ensemble weights for the matching systems various components. The *Learn* and *recommend* modes are a re-implementation of the system presented in [1] in which run-time complexity has been reduced from $O(n!)$ to $O(n^2)$ and generic interfaces have been provided to allow any matching system to use SMB by command-line invocation.

1.3 State, purpose, general statement

GeRoMeSuite is a generic system which can match ontologies as well as schemas in other modeling languages such as XML Schema or SQL. Therefore, it is well suited for matching tasks across heterogeneous modeling languages, such as matching XML Schema with OWL. We discussed in [6] that the use of a generic metamodel, which represents the semantics of the models to be matched in detail, is more advantageous for such heterogeneous matching tasks than a simple graph representation.

SMB is also a modeling language independent ‘meta’ matching system which mainly works on the similarity matrices produced by GeRoMeSuite. It improves the *clarity* of the similarity values by improving ‘good’ values and descreasing ‘bad’ values. This should increase the precision of the match result.

1.4 Specific techniques used

Besides the integration of GeRoMeSuite and SMB, we focused this year on adding validation methods to the system to improve the precision of the match result. A component for adding disjointness relationships in an ontology has been added to the matching framework. The component uses machine learning techniques to identify disjoint concepts with one ontology. The disjointness relationships can then be used in the validation of schema matches using logical reasoning.

Furthermore, we developed a component which can use a background ontology to find additional matches in the ontology. The system is able to find an appropriate background ontology on the web automatically, using Google and Swoogle. Due to the set up of the OAEI campaign, we did not use this component for OAEI.

1.5 Adaptations made for the evaluation

We evaluated several match configurations which is easily possible due to the adaptable and extensible matching framework of GeRoMeSuite. As only one configuration can be used for all matching tasks, we had to find a good compromise between performance in terms of precision and recall, time performance for larger ontologies (e.g., anatomy), and selection of appropriate matchers which work well on all tracks. For example, we also tested configurations which had an f-measure that was about 5% higher than the configuration which we eventually used, but these configurations did not work well on all tracks. The identification of good match configurations is a topic for future research.

Fig. 1 indicates the strategy which we used for the matching tasks in the benchmark track. All aggregation and filter steps use variable weights and thresholds, which are based on the statistical values of the input similarities.

The role matcher is a special matcher which compares the roles of model elements in our generic role-based metamodel. In principle, this results in matching only elements of the same type, e.g., classes with classes only and properties with properties only.

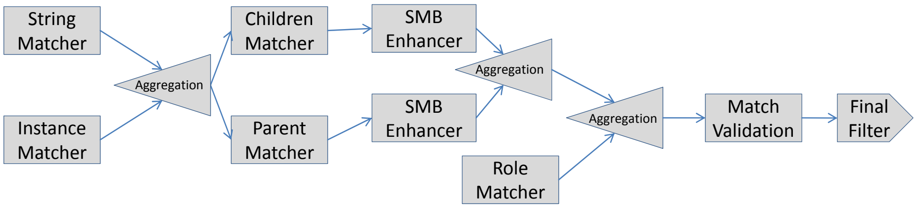

Fig. 1. Matching Strategy for OAEI 2010

On a technical level, we implemented a command line interface for the matching component, as the matching component is normally used from within the GUI of *GeRoMeSuite*. The command line interface can work in a batch mode in which several matching tasks and configurations can be processed and compared. The existence of this tool enabled also an easy integration with the OAEI web service interface.

1.6 Link to the system and parameters file

More information about the system can be found on the homepage of *GeRoMeSuite*: `http://www.dbis.rwth-aachen.de/gerome/oaei2010/`

The page provides also links to the configuration files used for the evaluation.

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

The results for the OAEI campaign 2010 are available at `http://www.dbis.rwth-aachen.de/gerome/oaei2010/`

2 Results

2.1 Benchmark

The following table shows the average results for precision and recall in the benchmark track.

Task	Precision	Recall
1xx	1,00	1,00
2xx (xx<48)	0,96	0,88
2xx (xx>47)	0,89	0,51
3xx	0,79	0,38

A first check, whether a match configuration is suitable at all are the 1xx ontologies. A configuration should produce the perfect result for these tracks, which is the case for the configuration, we have finally chosen.

For the simpler tasks in the 2xx data set (201-247), our system was able to achieve a very good result with an f-measure of more than 0.9.

For some of the really difficult tasks (248-266), our system was not able to find any correspondence as there is hardly any information that can be used (e.g., task 265 with no labels, no comments, no hierarchy, etc.).

The results for the tasks 3xx was in general good (f-measure of about 0.6 for 301, 302, and 304). However, ontology 303 is difficult for our generic system as the namespaces are not defined in a standard way. Therefore, we could only find a few correspondences.

2.2 Conference

The ontologies in the conference track are rather small and the matching tasks are more difficult as the ontologies have been designed by humans using different terminologies and having different goals in mind. As this is a more realistic case than the benchmark track, we have chosen a configuration which produces good results for the conference track. Using validation rules to check the logical consistency of the identified correspondences and a final filter step which generates only 1:1 correspondences was beneficial for the quality of the result.

At the current point, we can only report the results with respect to the reference alignments which are available. For these tasks, we achieve an average f-measure of about 0.45.

2.3 Anatomy

We participated in this task in the sub-tracks 1 to 3. Probably because of our validation and filtering methods, we achieved a high precision but low recall in task 1. Therefore, we used the result of task 1 also for task 2. In task 3, we achieved a high recall with respect to the *partial* reference alignment. We have to wait for the results with respect to the full alignment to make a final statement about the quality for this subtask.

2.4 Directory

We participate only in the single task modality of the directory track. The size of the input ontologies is similar to the anatomy track, so the same problems of scalability have to be faced here. We submitted an alignment with about 700 correspondences. Due to a missing reference alignment for the single task modality, we could not evaluate the quality of this result.

The main reason for not participating in the small task modality is that the small ontologies do not contain enough information to do a reasonable matching. Furthermore, we think that many of the given reference alignments are not correct.

3 Comments

We participate this time the third time in OAEI and see again some improvement of our matcher compared to last year. Thus, a structured evaluation and comparison of ontology alignment and schema matching components as OAEI is very useful for the development of such technologies. We appreciate especially the automatic evaluation system, although we also had to put some additional effort to get the interface and our web service working.

However, some reference alignments, especially in the directory track, should be reconsidered as they do not seem to be right. Furthermore, an oriented track as in OAEI 2009 would be useful to evaluate semantic matching techniques.

We are currently working on a system to generate a matching benchmark which comes closer to the challenges of real ontologies. We would be happy if we could contribute the results to OAEI 2011.

4 Conclusion

As our tool is neither specialized on ontologies nor limited to the matching task, we did not expect to deliver the best results. However, we are very satisfied with the overall results, as we can compete with the special purpose ontology alignment tools.

We will continue to work on the improvement of our matching system and on the integration of GeRoMeSuite and SMB. We will especially focus on the problem of identifying good match configurations automatically. We hope to participate again with an improved system in the OAEI campaign next year.

Acknowledgements: This work is supported by the DFG Research Cluster on Ultra High-Speed Mobile Information and Communication (UMIC, http://www.umic.rwth-aachen.de) and by the Umbrella Cooperation Programme (http://www.umbrella-coop.org/).

References

1. A. Gal, T. Sagi. Tuning the Ensemble Selection Process of Schema Matchers. *Information Systems*, **35**(8):845–859, 2010.
2. D. Kensche, C. Quix, M. A. Chatti, M. Jarke. *GeRoMe*: A Generic Role Based Metamodel for Model Management. *Journal on Data Semantics*, **VIII**:82–117, 2007.
3. D. Kensche, C. Quix, X. Li, Y. Li. *GeRoMeSuite*: A System for Holistic Generic Model Management. C. Koch, J. Gehrke, M. N. Garofalakis, D. Srivastava, K. Aberer, A. Deshpande, D. Florescu, C. Y. Chan, V. Ganti, C.-C. Kanne, W. Klas, E. J. Neuhold (eds.), *Proceedings 33rd Intl. Conf. on Very Large Data Bases (VLDB)*, pp. 1322–1325. Vienna, Austria, 2007.
4. D. Kensche, C. Quix, Y. Li, M. Jarke. Generic Schema Mappings. *Proc. 26th Intl. Conf. on Conceptual Modeling (ER'07)*, pp. 132–148. 2007.

5. C. Quix, D. Kensche, X. Li. Generic Schema Merging. J. Krogstie, A. Opdahl, G. Sindre (eds.), *Proc. 19th Intl. Conf. on Advanced Information Systems Engineering (CAiSE'07)*, *LNCS*, vol. 4495, pp. 127–141. Springer-Verlag, 2007.
6. C. Quix, D. Kensche, X. Li. Matching of Ontologies with XML Schemas using a Generic Metamodel. *Proc. Intl. Conf. Ontologies, DataBases, and Applications of Semantics (ODBASE)*, pp. 1081–1098. 2007.