Results of GeRoMeSuite for OAEI 2008

Christoph Quix, Sandra Geisler, David Kensche, Xiang Li

Informatik 5 (Information Systems)
RWTH Aachen University, Germany
`http://www.dbis.rwth-aachen.de`

Abstract. *GeRoMeSuite* is a generic model management system which provides several functions for managing complex data models, such as schema integration, definition and execution of schema mappings, model transformation, and matching. The system uses the generic metamodel *GeRoMe* for representing models, and because of this, it is able to deal with models in various modeling languages such as XML Schema, OWL, ER, and relational schemas.
A component for schema matching and ontology alignment is also part of the system. We participated this year the first time in the OAEI contest in order to evaluate and compare the performance of our matcher component with other systems. Therefore, we focused our efforts on the 'benchmark' track.

1 Presentation of the system

Manipulation of models and mappings is a common task in the design and development of information systems. Research in Model Management aims at supporting these tasks by providing a set of operators to manipulate models and mappings. As a framework, *GeRoMeSuite* [4] provides an environment to simplify the implementation of model management operators. *GeRoMeSuite* is based on the generic role based metamodel *GeRoMe* [3], 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 [7], Merge [6], and Compose [5].

The matching component of *GeRoMeSuite* has been described in more detail in [7], where we present and discuss in particular the results for heterogeneous matching tasks (e.g. matching XML Schema and OWL ontologies).

1.1 State, purpose, general statement

As a generic model management tool, *GeRoMeSuite* provides several matchers which can be used for matching models in general, i.e. our tool is not restricted to a particular domain or modeling language. Therefore, the tool provides several well known matching strategies, such as string matchers, Similarity Flooding, children and parent matchers, matchers using WordNet, etc. In order to enable the flexible combination of these basic matching technologies, matching strategies combining several matchers can be configured in a graphical user interface.

Because of its generic approach, *GeRoMeSuite* is well suited for matching tasks across heterogeneous modeling languages, such as matching XML Schema with OWL. We discussed in [7] 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.

Furthermore, *GeRoMeSuite* is a holistic model management and not limited to schema matching or ontology alignment. It supports also other model management tasks such as schema integration [6], model transformation [2], mapping execution and composition [5].

1.2 Specific techniques used

The basis of *GeRoMeSuite* is the representation of models (including ontologies) in the generic metamodel *GeRoMe*. Any kind of model is transformed first into the generic representation, then the model management operators can be applied to the generic representation. The main advantage of this approach is that operators have to be implemented only once for the generic representation. In contrast to other (matching) approaches which use a graph representation without detailed semantics, our approach is based on the semantically rich metamodel *GeRoMe* which is able to represent modeling features in detail.

For the OAEI campaign, we focused on improving our matchers for the special case of ontology alignment, e.g. we added some features which are useful for matching ontologies. For example, the generic representation of models allows the traversal of models in several different ways. During the tests with the OAEI tasks, we realized that, in contrast to other modeling languages, traversing the ontologies using another structure than class hierarchy is not beneficial. Therefore, we configured all our matchers that take the model structure into account just to work with the class hierarchy. Furthermore, we implemented so called ‘children’ and ‘parent’ matchers, which propagate the similarity of elements up and down in the class hierarchy.

In addition, we also implemented a matcher using WordNet to discover synonyms in the ontologies. However, as the benchmark track contains only one example which uses synonyms, we did not include this matcher in the final configuration for the OAEI campaign.

1.3 Adaptations made for the evaluation

As only one configuration can be used for all matching tasks, we worked on strategies for measuring the quality of an alignment without having a reference alignment. We compared several statistical measures (such as expected value, variance, etc.) of alignments with different qualities in order to identify a ‘good’ alignment. Furthermore, these values can be used to set thresholds automatically.

During the tests, we made the experience that the expected value of all similarities, the standard deviation, and the number of mappings per model element can be used to evaluate the quality of an alignment.

Fig. 1 indicates the strategy which we used for the matching tasks in the benchmark track. The aggregation and filter steps are not fixed, they use the statistical values of the input similarities to adapt the actual values of, for example, threshold and aggregation weights.

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 that only elements of the same type are matched, e.g. classes with classes only and properties with properties only.

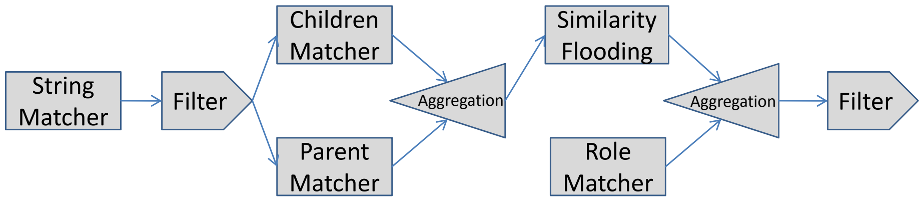


Fig. 1. Matching Strategy for OAEI

Furthermore, we experimented with histograms, i.e. a graphical representation of the distribution of similarity values. Although we could not identify particular patterns for histograms of ‘good’ or ‘bad’ alignments, we found the histogram quite useful for working interactively with the matching component of *GeRoMeSuite*. Fig. 2 shows a screenshot of *GeRoMeSuite*, including a window showing the histogram of a match result. In addition, the upper part of the window shows some statistical values for the current similarities. In another dialog, the filter can be adapted.

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 framework of *GeRoMeSuite*. The command line interface can work in a batch modus in which several matching tasks and configurations can be processed and compared.

1.4 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/`

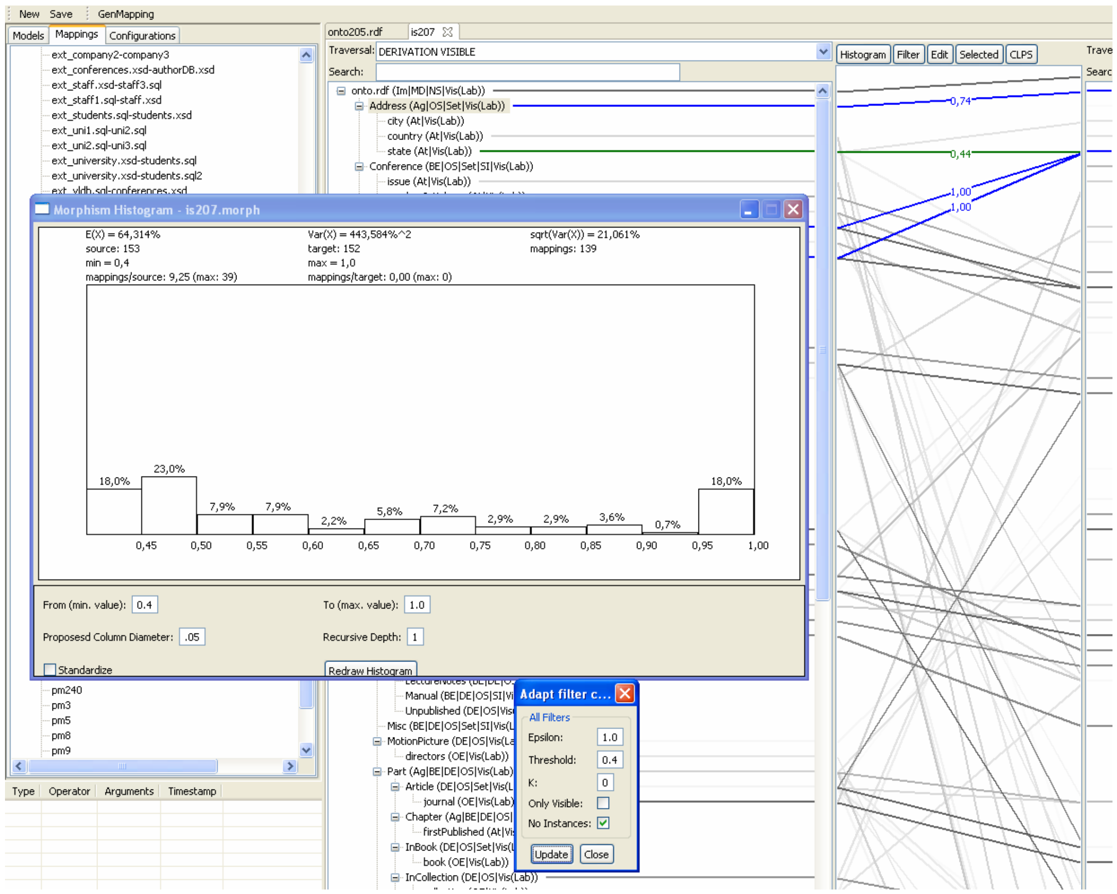
The page provides also links to the configuration files used for the evaluation.

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

The results for the OAEI campaign 2008 are available at `http://www.dbis.rwth-aachen.de/gerome/results.html`

2 Results

As we participated the first time in the OAEI campaign, we just focused on the benchmark track. The time used for each matching task was about 5 to 15 seconds.


New Save GenMapping
Models Mappings Configurations
ext_company2-company3
ext_conferences.xsd-authorDB.xsd
ext_staff.xsd-staff3.sql
ext_staff1.sql-staff.xsd
ext_students.sql-students.xsd
ext_uni1.sql-uni2.sql
ext_uni2.sql-uni3.sql
ext_university.xsd-students.sql
ext_university.xsd-students.sql2
pm240
pm3
pm5
pm8
pm9
Type Operator Arguments Timestamp
onto205.rdf is207
Traversal: DERIVATION VISIBLE
Search:
Histogram Filter Edit Selected CLPS
onto.rdf (Im|MD|NS|Vis(Lab))
Address (Ag|OS|Set|Vis(Lab))
city (At|Vis(Lab))
country (At|Vis(Lab))
state (At|Vis(Lab))
Conference (BE|OS|Set|SI|Vis(Lab))
issue (At|Vis(Lab))
Manual (BE|DE|OS|SI|Vi
Unpublished (DE|OS|Vis
Misc (BE|DE|OS|Set|SI|Vis(L
MotionPicture (DE|OS|Vis(La
directors (OE|Vis(Lab))
Part (Ag|BE|DE|OS|Vis(Lab)
Article (DE|OS|Set|Vis(L
journal (OE|Vis(Lab)
Chapter (Ag|BE|DE|OS|
firstPublished (At|Vi
InBook (DE|OS|Set|Vis(L
book (OE|Vis(Lab))
InCollection (DE|OS|Vis(Lab))
0,74
0,44
1,00
1,00
Morphism Histogram - is207.morph

E(X) = 64,314% Var(X) = 443,584%^2 sqrt(Var(X)) = 21,061%
source: 153 target: 152 mappings: 139
min = 0,4 max = 1,0
mappings/source: 9,25 (max: 39) mappings/target: 0,00 (max: 0)

From (min. value): 0.4 To (max. value): 1.0
Proposesd Column Diameter: .05 Recursive Depth: 1
Standardize
Redraw Histogram

Adapt filter c...
All Filters
Epsilon: 1.0
Threshold: 0.4
K: 0
Only Visible: ☐
No Instances: ✓
Update Close

Fig. 2. GUI of the Matching Component of *GeRoMeSuite* with Histogram Dialog

2.1 Benchmark

Overall, our matching component achieved very similar values for precision and recall, which seems to be rather unusual, if we compare our results with the results of other systems for previous years, where the precision was usually higher than recall.

Tasks 101-104 These tasks were quite easy as we could achieve a high precision and recall already with simple string matchers.

Task	Precision	Recall
101	1,00	0,79
103	0,94	0,79
104	0,95	0,79

For task 102 (irrelevant ontology), our matcher identified a few corresponding elements, such as year and yearValue, or date and year. Depending on the application of the mapping, such correspondences might be reasonable (e.g. for ontology merging).

Tasks 201-210 In these tasks, the linguistic information could not always be used as labels or comments were missing. After including also comments into the matching process, we could improve the match quality for these tasks significantly. For the synonym task (205), we also tested a matcher which uses WordNet to detect synonyms. However, as this matcher did not significantly improve the quality of the match result and required about three times more time than all other matchers together, we dropped the WordNet matcher from the final configuration.

Overall, the results are satisfying, except for the case 202, where no linguistic information at all was available.

Task	Precision	Recall
201	0,87	0,79
201-2	0,95	0,79
201-4	0,93	0,79
201-6	0,97	0,79
201-8	0,91	0,79
202	0,17	0,06
202-2	0,74	0,77
202-4	0,93	0,67
202-6	0,94	0,60
202-8	0,77	0,48
203	1,00	0,79
204	1,00	0,79
205	1,00	0,79
206	0,93	0,78
207	0,93	0,78
208	0,99	0,77
209	0,58	0,45
210	0,61	0,73

2.2 Tasks 221-231

The ontologies in these tasks lacked some structural information. As our matcher still uses string similarity in a first step, the results in this section were still quite reasonable.

Task	Precision	Recall
221	1,00	0,65
222	0,97	0,78
223	0,85	0,78
224	1,00	0,79
225	1,00	0,79
228	1,00	0,88
230	0,97	0,85
231	1,00	0,79

Tasks 232-266 These tasks are some combinations of the tasks before. For most of the tasks, the performance of our matcher was satisfying, but for some tasks, especially those without any linguistic information, it produced disappointing results. This gives some hints for future improvements of our matcher component, e.g. taking into account the overall structure of the ontology.

Tasks 301-304 For tasks 301 and 304, our system produce quite reasonable results. Further improvements could have been achieved, for example, by using the WordNet matcher for detecting synonyms, but we did not include this matcher because of performance reasons as explained above. Task 303 could not be processed by our system as there was a problem with importing this ontology into our generic representation.

3 Comments

A structured evaluation and comparison of ontology alignment and schema matching components is very useful for the development of such technologies. However, mappings between models are constructed for various reasons which can result in very different mapping results. For example, mappings for schema integration may differ from mappings for data translation. Therefore, different semantics for ontology alignments should be taken into account in the future, as it has been pointed out for schema matching in [1].

4 Conclusion

As our tool is neither specialized on ontologies nor limited to the matching task, we did not expect to deliver very good results. However, we were quite satisfied with the overall results. In general, we need to work on an improvement of the recall value. Furthermore, techniques used by other tools presented at the workshop would help us to improve the quality of the matching result and the performance of our tool. For example, identification of similar sub-structures in ontologies and semantic verification of the identified correspondences seem to be promising techniques to improve the quality and performance of the matching system.

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`).

References

1. J. Evermann. Theories of Meaning in Schema Matching: A Review. *Journal of Database Management*, **19**(3):55–82, 2008.
2. D. Kensche, C. Quix. Transformation of Models in(to) a Generic Metamodel. *Proc. BTW Workshop on Model and Metadata Management*, pp. 4–15. 2007.
3. 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.
4. 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.
5. D. Kensche, C. Quix, Y. Li, M. Jarke. Generic Schema Mappings. *Proc. 26th Intl. Conf. on Conceptual Modeling (ER'07)*, pp. 132–148. 2007.
6. 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, pp. 127–141. Springer-Verlag, 2007.
7. 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)*. 2007.