MOMA Framework - How to find a suitable matching approach

Malgorzata Mochol

Freie Universität Berlin, Institut für Informatik
Takustr. 9, D-14195 Berlin, Germany
mochol@inf.fu-berlin.de

Abstract. Current matching algorithms cannot be optimally used in ontology matching tasks as envisioned by the Semantic Web community, mainly because of the inherent dependency between approaches and ontology properties. As one possible solution we propose a Metadata-based Ontology MAtching (MOMA) framework based on a reuse-paradigm that, given a set of ontologies to be matched, takes into account the capabilities of existing matching algorithms and suggests appropriate ones for application. A Web-enabled matching framework requires a highly flexible selection and composition of available matching services in order to take full advantage of the broad spectrum of ontologies across the network. This will be reached by applying a flexible tool for decision-making and dependency rules-statements.

1 Motivation

"Would you be the first passenger on an airliner whose parts have just come out of the R&D shop? Or would you prefer to board knowing the aircraft was designed and constructed with parts that have successfully kept planes airborne for years? Practicing engineers, whether they specialize in civil, mechanical, electrical or aerospace engineering, select from components whose characteristics have been tested and proven for safety and efficiency." [22]. Reusability and reuse are fundamental principles of engineering, and their importance is also well accepted by the Semantic Web community. While ontology reuse has already been addressed in some methodologies [14, 15, 19, 21], the reuse of methodologies to merge, map or match ontologies has not yet been treated in depth within the community.

A fully developed Semantic Web will contain numerous distributed and ubiquitously available ontologies which will be used in different tasks and disciplines. A fundamental requirement for the realization of this vision is, first of all, proved and tested ontology matching algorithms (matchers), which are able to deal with the heterogeneity of ontological sources available on the Web. When implementing an application using a matching approach, the corresponding algorithm is typically built from scratch and does not resort to available methods. Despite the impressive number of research initiatives in the matching field containing valuable ideas and techniques, current matching approaches still feature important limitations when applied to the emerging Semantic Web: They require specific representation and natural languages [2], do not perform well on inputs with heterogeneous (graph) structures (e.g. Cupid [6]), or are restricted to tree-based conceptual models (e.g. S-Match [6]). Even combining multiple matching algorithms still does not overcome all the deficiencies: The quality of what is termed the meta-learner approaches (e.g. LSD [3]) is directly proportional to the efforts invested in training the algorithms [20], while black box solutions such as COMA [2] have proven to be unable to adapt well to complex application scenarios due to the inflexibility of the built-in matching combinations.

To make full use of the existing matching approaches, we propose a solution based on the reuse of approved algorithms. We argue that a possible solution to the matching dilemma is to design a new matching strategy – as opposed to a new (combined) matching heuristic – which strives to optimize the matching process (by reusing existing approaches) whilst being aware of the inherent dependencies between matchers, their execution characteristic, required output, and the types of ontologies they are able to process successfully.

2 General approach of the thesis

This thesis proposal analyzes the difficulties of current ontology matching approaches, with the aim of developing a new strategy based on software reuse paradigm to satisfy the core requirements of Semantic Web-compatible matching method and able to meet the demands of different users. We have been developing a matching framework (cf. Sec. 3) which serves:

• *human-matching users* (e.g. ontology engineers) who, for instance, wish to match two (or more) source ontologies to create a new target ontology based on a combination of existing knowledge resources; this process of choosing the suitable (w.r.t the given requirements) approach *can occur manually*;

• *machine-matching users* (e.g. service/matching providers, search engines looking for similar ontologies) that play a core role in enabling mediating Web Service interactions; in this case the process of choosing a suitable matching approach *must occur automatically*.

In reference to the above-mentioned goals, we defined our main problem as a question: *"How does one (human/machine) find a suitable matching approach w.r.t. given requirements?"* To find an answer to this question, at first we had to find possible factors that have an impact on (the selection of) an adequate matching approach. Accounting for the empirical findings of different case studies in ontology engineering, regarding the requirements collected during the development of different Semantic Web application scenarios[1], and in intensive collaboration with ontology and software engineers, we defined six groups of factors (*dimensions*) as relevant for the matching selection process: (i)input (ontologies) characteristic; (ii)output characteristic; (iii)approach (matching algorithms) attributes; (iv)usability of the approach; (v)documentation of the approach and (vi)costs of the approach. In accordance with established ontology engineering

[1] Projects: *Wissensnetze*, *Reisewissen*, *SWPatho* `http:\\[wissensnetze|reisewissen|swpatho].ag-nbi.de`

methodologies [5] and, additionally, by analyzing the common research in the matching field [1, 4], we then refined the dimensions into factors (e.g. factors: *input size* and *input formality level* for dimension *input characteristic*) and further refined the factors into attributes (e.g. attributes: *number of ontologies* and *number of specific ontological primitives* for factor *input size*). These dimensions, factor and attributes can be expressed as a simple form (questionnaire) which can be understood by the human-matching users. However, to ensure a rich and at the same time formal representation of the ambiguous semantics of these factors, and to enable their integration and exchange in and between Semantic Web applications, we also modelled this information in ontological form and implemented it using Semantic Web representation languages. The chosen dimensions, factor and attributes building the metadata ((i): ontology metadata and (ii)-(vi): matching metadata, cf. Sec 3.1) do not stand alone but are rather in close relationship to each other. For this reason, we need powerful and flexible tools for decision-making in complex multi-criteria problems that will allow us, by linking these various characteristics, to determine manually - with user interaction - and automatically which matchers are to be used for the given input. Both approaches supporting manual as well as automated processes of choosing the appropriate matching form the so-called **M**etadata-based **O**ntology **MA**tching Framework - *MOMA framework* (cf. Sec. 3).

The particular MOMA components and the framework as a whole have to be evaluated. The evaluation of metadata was performed by means of an in-situ experiment, in which the resulting model is compared against the requirements derived from the different case studies. Since the framework supports two selection ways of appropriate matchings, we need two evaluation methods: the *MOMA-manual selection approach* supported by the Web-based tool is to be evaluated against manual selection without any tool support (tests will be conducted by different users looking for the suitable matchings w.r.t the same given requirements), and the *MOMA-automatic selection approach* is to be evaluated in the context of different case studies.

3 MOMA Framework

We foresee several usage patterns for the proposed MOMA framework. On one hand, data and service providers can systematically publish their resources, be they ontologies or new matchers, Web-wide with the help of the MOMA framework. In so doing, they are expected to provide the required descriptive metadata for the resources subscribed, as this guarantees a higher visibility of their products to the incoming inquiries. On the other hand, matching consumers (humans and machines) consult the MOMA framework to get (information about) matching algorithms adequate for their particular tasks.

In the following we describe the matching and ontology metadata and their role within the MOMA-selection process of suitable matchings.

3.1 MOMA Metadata

The MOMA matching metadata[2] captures information (*(ii)-(vi) dimensions*) about existing ontology matchers. For the classification of the algorithms, we are initially relying on Rahm and Bernstein [16], who make the distinction between *individual* and *combining* matchers[3], but in further developments the metadata will also take into account additional subsumptions and characteristics proposed e.g. in [12] and [18]. Beside the aforementioned classification, the metadata model includes different factors and attributes, such as input type (instances or schemas, additional input in the form of e.g. numerical values), matching level (atomic level, e.g. attributes in an XML schema and non-atomic level e.g. XML elements), and cardinality (i.e. whether a matcher compares one or more elements in one schema with one or more elements in a second schema).

The dimension *output characteristic* including factors and attributes (MOMA ontology metadata) is described using the metadata model[4] introduced in [13], which can be used to describe ontologies in various phases of their life-cycle. Accounting for the fact that matching algorithms cannot be applied with equal success regardless of any parts of the ontology metadata model, we have identified the following ontology factors as relevant for matching tasks:

- *syntactic features*, such as *number of specific ontological primitives* that affect the matching execution performance and quality of, for example, structured-based matchers that typically perform better on simple graph structures;
- *semantic features*, such as modelled domain, representation/natural language, level of formality or domain generality that restrict the number of applicable matching algorithms, which might be adequate for a sub-set of these features.

The MOMA metadata models are defined in a formal representation language (currently OWL) and the dependencies between the *dimensions* are defined on the basis of a decision-making approach or formalized in terms of rules.

3.2 Selection of the suitable matcher

For a given pair of ontologies to be matched, the MOMA framework must decide which matching algorithms are to be applied to satisfy the requirements and to obtain the desired output. To support a manual selection of the suitable matching approaches we decided to adopt Analytic Hierarchy Process (AHP) [17] — an approach that had been already successfully applied for different selections especially for reuse decisions [9, 10]. AHP allows the gathering of knowledge about a particular problem, and the quantification of subjective opinions as well to forcing the comparison of alternatives in relation to established criteria. By reducing complex decisions (which matching is suitable) to a series of pair-wise comparisons (dimensions, factors and attributes) and synthesizing the results

[2] Available at `http://wissensnetze.ag-nbi.de/matching/matchingmetadata.owl`
[3] Detailed description in [11]
[4] Available at `http://swpatho.ag-nbi.de/context/meta.owl` and `http://omv.ontoware.org`.

(list of suitable algorithms), decision-makers arrive at the best decision based on a clear rationale [17].

In order to *automatically* determine which matchers are to be used (or excluded) we propose to formalize the knowledge concerning dependencies between these algorithms and the inputs on which they operate in terms of dependency rules-statements (e.g. if no instance data is available, then apply only scheme matchers; apply only matchers which are capable of dealing with the representation language of the inputs). For the purpose of the first implementation, the matching rules are put into effect in SWRL [8], a rule language for the Semantic Web which the rules formalized in terms of the concepts defined in the two metadata models. However, the deployment of rules in decision-making processes requires a reasoning engine able to operate on ontologies and (SWRL) rules, an issue which is still the subject of active research in the Semantic Web community.

4 Summary and Future work

This paper presents ideas of the first two years of work on the proposed PhD thesis which, taking into account problems and limitations of both individual and combining matchers, aims to develop an ontology matching framework relayed on a reuse-strategy of existing algorithms. Our strategy is based on dependencies between algorithms and the types of ontologies these are able to process successfully, and factors that have influence on the matching tasks. Since two different policies, the manual, based on AHP, and the automated, applying rules, are to be integrated into the MOMA framework, it can serve both human as well as machine users.

The preliminary version of the prototypical implementation of the MOMA framework contains, in consideration of dimensions, factors and attributes, the ontology metadata and (the first version of) matching metadata, as well a few general prescribed rules. The first phase of further development will be dedicated to expansion of the matching metadata, the application of AHP for the manual decision-making process, and the definition of the comprehensive set of rules using information related to the factors of matchings and ontologies for the automated detection of suitable matchers.

Acknowledgements This work has been partially supported by the "Knowledge Nets" project, which is part of the InterVal - Berlin Research Centre for the Internet Economy, funded by the German Ministry of Research BMBF and by the KnowledgeWeb - Network of Excellence. Further information about MOMA Framework can be found under `http://moma.ag-nbi.de`.

References

1. P. et al. Bouquet. D2.2.1 Specification of a common framework for characterizing alignment. Technical Report KWEB/2004/D2.2.1/v2.0, KnowledgeWeb Network of Excellence, `http://www.inrialpes.fr/exmo/cooperation/kweb/heterogeneity/deli/kweb-2%21.pdf`, February 2005.

2. H. H. Do and E. Rahm. COMA—a system for flexible combination of schema matching approaches. In *Proc. of the 28th VLDB Conference*, 2002.
3. A. Doan, P. Domingos, and A. Halevy. Reconciling Schemas of disparate Data sources: A Machine Learning Approach. In *Proc. of the SIGMOD Conference*, 2001.
4. J. Euzenat, R. Garca Castro, and M. Ehrig. D2.2.2 Specification of a benchmarking methodology for alignment techniques. Technical Report KWEB/2004/D2.2.2/v1.0, Knowledge Web Networkl of Excellence, February 2005.
5. M. Fernandez-Lopez and A. Gomez-Perez. Overview and analysis of methodologies for building ontologies. *Knowledge Engineering Review*, 17(2):129–156, 2002.
6. F. Giuchiglia and P. Shvaiko. Semantic Matching. *Knowledge Web Review Journal*, pages 265–280, 2004.
7. J. Hartmann, E. Paslaru Bontas, R. Palma, and A. Gomez-Perez. DEMO - A Design Environment for Metadata About Ontologies. In *Proc. of the 3rd European Semantic Web Conference ESWC2006 (to be published)*, 2006.
8. I. Horrocks, P. F. Patel-Schneider, H. Boley, S. Tabet, B. Grosof, and M. Dean. SWRL: A Semantic Web Rule Language Combining OWL and RuleML. `http://www.w3.org/Submission/SWRL`, 2004.
9. J. Kontio. A Case Study in Applying a Systematic Method for COTS Selection. In *Proc. of 18th International Conference on Software Engineering*, 1996.
10. A. Lozano-Tello and A. Gomez-Perez. ONTOMETRIC: A Method to Choose the Appropriate Ontology. *Journal of Database Management*, 15:1–18, 2004.
11. M. Mochol and E. Paslaru Bontas. A Metadata-Based Generic Matching Framework for Web Ontologies. Technical Report TR-B-05-03, FU Berlin, `http://www.inf.fu-berlin.de/inst/pubs/tr-b-05-08.abstract.html`, June 2005.
12. P. Mork and P. A. Bernstein. Adapting a Generic Match Algorithm to Align Ontologies of Human Anatomy. In *Proc. of the 20th International Conference on Data Engineering (ICDE'04)*, page 787, 2004.
13. E. Paslaru Bontas. Using Context Information to Improve Ontology Reuse. In *Proc. of the Doctoral Consortium at the CAISE'05*, 2005.
14. E. Paslaru Bontas and M. Mochol. Towards a reuse-oriented methodology for ontology engineering. In *Proc. of 7th International Conference on Terminology and Knowledge Engineering (TKE 2005)*, 2005.
15. B. J. Peterson, W. A. Andersen, and J. Engel. Knowledge Bus: Generating Application-focused Databases from Large Ontologies. In *Proc. of the KRDB'98*, 1998.
16. E. Rahm and P. A. Bernstein. A survey of approaches to automatic schema matching. *Journal of Very Large Data Bases*, 2001.
17. T. L. Saatly. How to Make a Decision: The Analytic Hierarchy Process. *European Journal of Operational Research*, (48):9–26, 1990.
18. P. Shvaiko. Iterative schema-besed semantic matching. Technical Report DIT-04-020, University of Trento, `http://eprints.biblio.unitn.it/archive/00000550/01/020.pdf`, June 2004.
19. B. Swartout, R. Patil, K. Knight, and T. Russ. Toward Distributed Use of Large-Scale Ontologies. In *Proc. of the Tenth Knowledge Acquisition for Knowledge-Based Systems Workshop*, 1996.
20. K. Tu and Y. Yu. CMC: Combining Multiple Schema-Matching Strategies based on Credibility Prediction. In *Proc. of 10th International Conference on Database Systems for Advanced Applications (DASFAA 2005)*, 2005.

21. M. Uschold and M. King. Towards a Methodology for Building Ontologies. In *Proc. of the IJCAI'95, Workshop on Basic Ontological Issues in Knowledge Sharing*, 1995.
22. M. Williamson. Special Section: Software Reuse. *CIO Magazine*, 3 1997.