

OntoBuilder: Fully Automatic Extraction and Consolidation of Ontologies from Web Sources

Avigdor Gal
Technion - Israel Institute of Technology
avigal@ie.technion.ac.il

Giovanni Modica
Mississippi State University
modicag@hotmail.com

Hasan Jamil
Wayne State University
jamil@cs.wayne.edu

Ontologies, formal specifications of domains, have evolved in recent years as a leading tool in representing and interpreting Web data. The inherent heterogeneity of Web resources, the vast amount of information on the Web, and its non-specific nature requires a semantically rich tool for extracting the essence of Web sources' content. The OntoBuilder project [3, 2] supports the extraction of ontologies from Web search interfaces, ranging from simple Search Engine forms to multiple-pages, complex reservation systems. Ontologies from similar domains are then consolidated into an ever improving single ontology with which a domain can be queried, either automatically or semi-automatically.

Given a sample form, filled by the user, and given a new form, from another Web site, OntoBuilder finds the best mapping between the two forms. This, in turn, can serve a system in automatically filling the fields (a sort of a query rewriting), according to the mapping suggested by OntoBuilder.

OntoBuilder enables fully-automatic ontology matching. The use of ontologies, as opposed to relational schema or XML, as an underlying data model allows a flexible representation of metadata, that can be tailored to many different types of applications. OntoBuilder contains several unique matching algorithms that can match concepts (terms) by their data types, constraints on value assignment, and above all, the ordering of concepts within forms (termed *precedence*).

OntoBuilder was developed using Java, which makes it portable to various platforms and operating system environments. OntoBuilder generates dictionary of terms by extracting labels and field names from Web forms, and then it recognizes unique relationships among terms, and utilize them in its matching algorithms.

OntoBuilder is a generic tool and serves as a module for several projects. For example, we have designed a frame-

work for evaluating automatic schema matching algorithms [1], and we use OntoBuilder both for evaluation and for improving our methodology. This framework provides a sufficient condition (we term *monotonicity*) for a matching algorithm to generate "good" ontologies. Our empirical results with OntoBuilder show that its algorithms satisfy one of the forms of monotonicity we present in [1]. Also, algorithms from OntoBuilder are being employed in an agent negotiation protocol for trading information goods.

We demonstrate OntoBuilder using an easy-to-follow example of matching Car rental ontologies. The system will create ontologies of car rental Web sites on-the-fly, and combine them into a global ontology. The benefits of OntoBuilder in resolving, in an automatic manner, semantic heterogeneity, including synonyms and designer errors, will be highlighted.

OntoBuilder is available at
<http://ie.technion.ac.il/OntoBuilder>.

References

- [1] A. Gal, A. Anaby-Tavor, A. Trombetta, and D. Montesi. A framework for modeling and evaluating automatic semantic reconciliation. *VLDB Journal*, 2004. to appear.
- [2] A. Gal, G. Modica, H. Jamil, and A. Eyal. Automatic ontology matching using application semantics. Submitted for publication. Available upon request from avigal@ie.technion.ac.il, 2003.
- [3] G. Modica, A. Gal, and H. Jamil. The use of machine-generated ontologies in dynamic information seeking. In C. Batini, F. Giunchiglia, P. Giorgini, and M. Mecella, editors, *Cooperative Information Systems, 9th International Conference, CoopIS 2001, Trento, Italy, September 5-7, 2001, Proceedings*, volume 2172 of *Lecture Notes in Computer Science*, pages 433–448. Springer, 2001.