ONTOLOGY LEARNING FOR THE SEMANTIC WEB

Alexander Maedche

Kluwer Academic Publishers

ONTOLOGY LEARNING FOR THE SEMANTIC WEB

Alexander Maedche
University of Karlsruhe, Germany

KLUWER ACADEMIC PUBLISHERS
Boston / Dordrecht / London

Distributors for North, Central and South America:
Kluwer Academic Publishers
101 Philip Drive
Assinippi Park
Norwell, Massachusetts 02061 USA
Telephone (781) 871-6600
Fax (781) 871-6528
E-Mail <kluwer@wkap.com>

Distributors for all other countries:
Kluwer Academic Publishers Group
Distribution Centre
Post Office Box 322
3300 AH Dordrecht, THE NETHERLANDS
Telephone 31 78 6392 392
Fax 31 78 6546 474
E-Mail <services@wkap.nl>

Electronic Services <http://www.wkap.nl>

Library of Congress Cataloging-in-Publication Data

Maedche, Alexander D.
Ontology learning for the semantic Web / by Alexander D. Maedche.
p. cm.
Includes bibliographical references and index.
ISBN 0-7923-7656-0 (alk. paper)
1. Web site development. 2. Metadata. 3. Ontology. 4. Artificial intelligence. I. Title.

TK5105.888 .M33 2002
005.2'76—dc21

2001058188

Printed on acid-free paper.

Printed in the United States of America

ONTOLOGY LEARNING FOR THE SEMANTIC WEB

THE KLUWER INTERNATIONAL SERIES IN ENGINEERING AND COMPUTER SCIENCE

List of Tables

Preface

The web in its' current form is an impressive success with a growing number of users and information sources. However, the growing complexity of the web is not reflected in the current state of Web technology. The heavy burden of accessing, extracting, interpretating and maintaining is left to the human user. Tim Berners-Lee, the inventor of the WWW, coined the vision of a Semantic Web in which background knowledge on the meaning Web resources is stored through the use of machine-processable (meta-)data. The Semantic Web should bring structure to the content of Web pages, being an extension of the current Web, in which information is given a well-defined meaning. Thus, the Semantic Web will be able to support automated services based on these descriptions of semantics. These descriptions are seen as a key factor to finding a way out of the growing problems of traversing the expanding web space, where most web resources can currently only be found through syntactic matches (e.g., keyword search).

Ontologies have shown to be the right answer to these structuring and modeling problems by providing a formal conceptualization of a particular domain that is shared by a group of people. Thus, in the context of the Semantic Web, ontologies describe domain theories for the explicit representation of the semantics of the data. The Semantic Web relies heavily on these formal ontologies that structure underlying data enabling comprehensive and transportable machine understanding. Though ontology engineering tools have matured over the last decade, the manual building of ontologies still remains a tedious, cumbersome task which can easily result in a knowledge acquisition bottleneck. The success of the Semantic Web strongly depends on the proliferation of ontologies, which requires that the engineering of ontologies be completed quickly and easily. When using ontologies as a basis for Semantic Web applications, one has to face exactly this issue and in particular questions about development time, difficulty, confidence and the maintenance of ontologies. Thus, what one ends up with is similar to what knowledge engineers have dealt with over the last two decades when elaborating methodologies for knowledge acquisition or workbenches for defining knowledge bases. A method which has proven to be extremely beneficial for the knowledge acquisition task is the integration of knowledge acquisition with machine learning techniques.

This book is based on the idea of applying knowledge discovery to multiple data sources to support the task of developing and maintaining ontologies. The notion of Ontology Learning aims at the integration of a multitude of disciplines in order to facilitate the construction of ontologies, in particular machine learning. Ontology Learning greatly facilitates the construction of ontologies by the ontology engineer. The vision of Ontology Learning that is proposed here includes a number of complementary disciplines that feed on different types of unstructured and semi-structured data in order to support a semi-automatic ontology engineering process. Because the fully automatic acquisition of knowledge by machines remains in the distant future, the overall process is considered to be semi-automatic with human intervention. It relies on the "balanced cooperative modeling" paradigm, describing a coordinated interaction between human modeler and learning algorithm for the construction of ontologies for the Semantic Web. This objective in mind, an approach that combines ontology engineering with machine learning is described, feeding on the resources that we nowadays find on the Web.

This book is split into four parts: In the first part the basics on the history of ontologies, as well as their engineering and embedding into applications for the Semantic Web are systematically introduced. This portion of the book includes a formal definition of what an ontology is and a collection of ontology-based application examples in the Semantic Web. Subsequently, a layered ontology engineering framework is introduced. The framework uses a layered representation of ontologies based on W3C standards such as RDF(S) and its' current extensions being created by the knowledge engineering and representation community. The second part establishes a generic framework for Ontology Learning for the Semantic Web. It discusses a wide range of different types of existing data on the current Web relevant to Ontology Learning. The Ontology Learning framework proceeds through ontology import, extraction, pruning and refinement and gives the ontology engineer a wealth of coordinated tools for ontology engineering. Besides the general framework and architecture, a number of techniques for importing, processing and learning from existing data are introduced, such as HTML documents and dictionaries. The third part of the book describes the implementation and evaluation of the proposed ontology learning framework. First, it describes the developed ontology engineering workbench, ONTOEDIT, supporting manual engineering and the maintenance of ontologies based on the fundamentals introduced in the first part of the book. Second, the ontology learning environment TEXT-TO-ONTO implements the ontology learning framework as shown in the second chapter of the book. An important aspect of applying ontology learning techniques deals with the question of how to measure the quality of the application of these techniques. Therefore, the third part of this book introduces a new approach and measures for evaluating ontology learning based on the well-known idea of having gold standards as evaluation references. The fourth part of this book provides a detailed overview of existing work that emphasizes topics of interest with similarities to the task of ontology learning. It analyzes a multitude of disciplines (ranging from information retrieval, information extraction and machine learning to databases). The book concludes with a summary of contributions and insights gained. Finally, a vision of the future and a discussion of future challenges in regards to the Semantic Web is delineated.

ALEXANDER MAEDCHE

Acknowledgements

Writing a book is a complex project in that many people are involved. I thank all people supporting me in my research and especially in writing this book. I appreciate very much the important roles that my colleagues Michael Erdmann, Siegfried Handschuh, Andreas Hotho, Gerd Stumme, Nenad Stojanovic, Ljiljana Stojanovic, York Sure, and Raphael Volz played. I thank all my students that supported me in my work by doing implementation and evaluation work. Very special thanks to Raphael Volz, now one of my colleagues, who did heavy implementation work in his master thesis. Stefan Decker, the Semantic Web initiator at our research group in Karlsruhe, always and at any time was open for useful comments. Special thanks to Steffen Staab for giving me the first ideas on Ontology Learning for the Semantic Web. He always was open for crazy discussions producing new ideas. I thank Rudi Studer, my advisor and leader of the research group. He supported me in making great experiences during my time at Karlsruhe. His way of leading me and the overall research group created a prolific research environment. Thanks to Jörg-Uwe Kietz that provided useful input and comments to my work on ontology learning. Without all of them them, this work would not have been possible.

I thank may parents that financed and supported my long stay at the university. Mostly, however, I must thank my friend and wife, Ellen, who always accepted when I was saying that there will come better times with less work. Thank to all of you for being there.

Alexander Maedche
Karlsruhe, Germany

Foreword

The success of the Web today can be explained to a large extent by its simplicity, i.e. the low level technical know-how that is needed to put information into the Web and to access Web information by browsing and keyword-based search. However, the volume of information that is nowadays available on the Web makes the limits of the current Web drastically obvious for its users: finding relevant information among millions of Web pages becomes more and more a heavy burden, and more than once it becomes impossible.

The development of the Semantic Web is a promising path towards transforming the Web into a semantically grounded information space that makes information accessible in a semantic way. It is a common understanding that machine-processable metadata that come with a semantic foundation as provided by ontologies, establish the technological basis for such a semantic processing of Web information.

All experience in practical settings shows that the engineering of ontologies is a crucial bottleneck when setting up Semantic Web applications. Furthermore, in fast changing market environments outdated ontologies mean outdated applications. As a consequence, the systematic management of the evolution of ontologies is a bottleneck as well.

Rather recently, these challenges gave rise to a new research area: "Ontology Learning". Ontology Learning aims at developing methods and tools that reduce the manual effort for engineering and managing ontologies. Ontology Learning is an inherently interdisciplinary area bringing together methods from ontology engineering, knowledge representation, machine learning, computational linguistics and information extraction. Nowadays, there is no chance to fully automate these learning processes. Therefore all approaches assume some cooperation between humans and machines, i.e. they provide semi-automatic means for ontology engineering and evolution.

This book describes a comprehensive framework for Ontology Learning. This framework addresses for the first time the specific aspects of Ontology

Learning that arise in the context of the Semantic Web, e.g. the heterogeneity of the Web sources and the layered representation of Web-based ontologies.

Ontology Learning relies on a tight integration of shallow linguistic processing with ontology representation. Therefore, the Ontology Learning framework defines a new notion of ontology that establishes precisely defined links between a linguistic layer, an ontology, and an associated knowledge base that populates the ontology. This integration paves the way for transforming lexical entries and linguistic asssociations into conceptual entries of the ontology and related conceptual relations.

The framework exploits a process-oriented view for Ontology Learning that distinguishes between the phases Import, Extract, Prune, and Refine. Thus, Ontology Learning is decomposed into subtasks that address specific aspects and can therefore solved with methods that are tailored to these subtask-specific challenges. Given the heterogeneity of the sources that are available in the Web context as well as the diversity of the different ontology learning tasks it is obvious that no single learning approach can meet all these different requirements. Therefore, the framework defines a system architecture that supports multistrategy learning, i.e. the results of different learning methods are combined in order to achieve sufficiently good learning results. Thus, the framework is open for adding new learning algorithms that may improve the learning results. The description of the framework elaborates different learning subtasks, especially the import of ontologies (including ontology integration), the extraction of ontologies from semi-structured sources, the learning of non-taxonomic relations, and the pruning of ontologies. As such, a broad collection of techniques is integrated into the Ontology Learning framework. A considerable part of the framework have been implemented in the ontology engineering framework OntoEdit and the learning environment Text-To-Onto.

When learning ontologies an immediate question arises: what is the quality of the learning results. This is a rather tough problem since there do not exist obvious quality standards. The ontology learning framework addresses this problem by introducing a collection of measures for comparing ontologies to each other. First evaluations indicate that the manual engineering and the learning of ontologies supplement each other in a nice way and thus open the way for further elaborating of how to arrange the cooperation between human and machine for ontology learning.

The ontology learning framework as described in this book is a promising step in further developing the field of ontology learning. By identifying clearly defined subtasks, further learning methods may be developed that enhance the learning results for respective subtasks. The framework is part of the development and implementation of the Karlsruhe Ontology and Semantic Web infrastructure that provides an overall architecture for managing and applying ontologies in the context of the Semantic Web. Thus ontology learning is tightly integrated with other aspects of the Semantic Web, like e.g semi-automatic generation of metadata, the alignment of ontologies or inferring new facts from given metadata and ontologies.

Ontology learning is a rather young, yet very promising research field. The transfer of its research results into scalable products will be an important step towards making the Semantic Web happen.

R. Studer, University of Karlsruhe

Contents

Part II Ontology Learning for the Semantic Web