
**THE KLUWER INTERNATIONAL SERIES
IN ENGINEERING AND COMPUTER SCIENCE**

ONTOLOGY LEARNING FOR THE SEMANTIC WEB

ONTOLOGY LEARNING FOR THE SEMANTIC WEB

by

Alexander Maedche
University of Karlsruhe, Germany



SPRINGER SCIENCE+BUSINESS MEDIA, LLC

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 978-1-4613-5307-2 ISBN 978-1-4615-0925-7 (eBook)

DOI 10.1007/978-1-4615-0925-7

1. Web site development. 2. Metadata. 3. Ontology. 4. Artificial intelligence. I. Title.

TK5105.888 .M33 2002

005.2'76—dc21

2001058188

Copyright © 2002 by Springer Science+Business Media New York

Originally published by Kluwer Academic Publishers. in 2002

Softcover reprint of the hardcover 1st edition 2002

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, mechanical, photo-copying, recording, or otherwise, without the prior written permission of the publisher.

Printed on acid-free paper.

Contents

List of Figures	ix
List of Tables	xiii
Preface	xv
Acknowledgements	xviii
Foreword by R. Studer	xix
 Part I Fundamentals	
1. INTRODUCTION	3
1 Motivation & Problem Description	3
2 Research Questions	4
3 Reader's Guide	6
2. ONTOLOGY — DEFINITION & OVERVIEW	11
1 Ontologies for Communication – A Layered Approach	15
2 Development & Application of Ontologies	21
3 Conclusion	25
3. LAYERED ONTOLOGY ENGINEERING	29
1 Ontology Engineering Framework	30
2 Layered Representation	34
3 Conclusion	49
3.1 Further Topics in Ontology Engineering	50
3.2 Ontology Learning for Ontology Engineering	51

Part II Ontology Learning for the Semantic Web

4. ONTOLOGY LEARNING FRAMEWORK	59
1 A Taxonomy of Relevant Data for Ontology Learning	60
2 An Architecture for Ontology Learning	66
2.1 Overview of the Architecture Components	66
2.2 Ontology Engineering Workbench ONTOEDIT	68
2.3 Data Import & Processing Component	70
2.4 Algorithm Library	71
2.5 Graphical User Interface & Management Component	72
3 Phases of Ontology Learning	73
3.1 Import & Reuse	74
3.2 Extract	75
3.3 Prune	76
3.4 Refine	77
4 Conclusion	78
5. DATA IMPORT & PROCESSING	81
1 Importing & Processing Existing Ontologies	83
1.1 Ontology Wrapper & Import	84
1.2 FCA-MERGE — Bottom-Up Ontology Merging	85
2 Collecting, Importing & Processing Documents	95
2.1 Ontology-focused Document Crawling	95
2.2 Shallow Text Processing using SMES	97
2.3 Semi-Structured Document Wrapper	105
2.4 Transforming Data into Relational Structures	107
3 Conclusion	112
3.1 Language Processing for Ontology Learning	112
3.2 Ontology Learning from Web Documents	113
3.3 (Multi-)Relational Data	114
6. ONTOLOGY LEARNING ALGORITHMS	117
1 Algorithms for Ontology Extraction	118
1.1 Lexical Entry Extraction	118
1.2 Taxonomy Extraction	122
1.3 Non-Taxonomic Relation Extraction	130
2 Algorithms for Ontology Maintenance	140
2.1 Ontology Pruning	140
2.2 Ontology Refinement	142

3	Conclusion	144
3.1	Multi-Strategy Learning	145
3.2	Taxonomic vs. Non-Taxonomic Relations	145
3.3	A Note on Learning Axioms — $\mathcal{A}^O$	146
Part III Implementation & Evaluation		
7.	THE TEXT-TO-ONTO ENVIRONMENT	151
1	Component-based Architecture	153
2	Ontology Engineering Environment ONTOEDIT	154
3	Components for Ontology Learning	163
4	Conclusion	168
8.	EVALUATION	171
1	The Evaluation Approach	172
2	Ontology Comparison Measures	173
2.1	Precision and Recall	174
2.2	Lexical Comparison Level Measures	175
2.3	Conceptual Comparison Level Measures	177
3	Human Performance Evaluation	183
3.1	Ontology Engineering Evaluation Study	184
3.2	Human Evaluation – Precision and Recall	185
3.3	Human Evaluation – Lexical Comparison Level	187
3.4	Human Evaluation – Conceptual Comparison Level	188
4	Ontology Learning Performance Evaluation	190
4.1	The Evaluation Setting	191
4.2	Evaluation of Lexical Entry Extraction	191
4.3	Evaluation of Concept Hierarchy Extraction	193
4.4	Evaluation of Non-Taxonomic Relation Extraction	194
5	Conclusion	196
5.1	Application-oriented Evaluation	197
5.2	Standard Datasets for Evaluation	198

Part IV Related Work & Outlook

9. RELATED WORK	203
1 Related Work on Ontology Engineering	204
2 Related Work on Frameworks of KA & ML	209
3 Related Work on Data Import & Processing	212
4 Related Work on Algorithms	214
5 Related Work on Evaluation	219
10. CONCLUSION & OUTLOOK	223
1 Contributions	223
2 Insights into Ontology Learning	224
3 Unanswered Questions	225
4 Future Research	226
References	228
Index	242

List of Figures

1.1	Reading this Book	8
2.1	The Meaning Triangle	14
2.2	Ontologies for Communication	15
2.3	Example: Instantiated Ontology Structure	19
2.4	Different Kinds of Ontologies	22
2.5	Relational Metadata on the Semantic Web	24
3.1	Layered Ontology Engineering	31
3.2	Representation Layers	35
3.3	An RDF Example	36
3.4	An RDF-Schema Example	38
3.5	An Example for the RDF-Schema Serialization Syntax	40
3.6	XML Serialization of RDF Instances	40
3.7	OntoEdit Representation Vocabulary	41
3.8	A Concrete Representation of the Lexicon	42
3.9	OIL Extensions of RDF(S)	42
3.10	Extending RDF(S) using Semantic Patterns	46
4.1	Taxonomy of Relevant Data for Ontology Learning	60
4.2	Architecture for Ontology Learning	67
4.3	OntoEdit Screenshot	69
4.4	Ontology Learning Cycle	73
5.1	Import and Processing Modules	82
5.2	WordNet and GermaNet Example	86
5.3	Ontology Merging Method	88
5.4	Two Example Contexts K_1 and K_2	90
5.5	The Pruned Concept Lattice	91

5.6	Natural Language Processing System Architecture	98
5.7	Example SMES Output – Morphological Component	101
5.8	Dependency Grammer Description	102
5.9	Example SMES Output – Underspecified Dependency Structure (abbreviated)	103
5.10	Example for a Heuristic Concept Relation	105
5.11	Example Normalized Dictionary Entry in RDF	106
5.12	Concept Matrix Generation View	111
6.1	Hierarchy Clustering with Labeling	127
6.2	Example Pattern for Dictionary Descriptions	129
6.3	Dictionary-based Extracted Concept Hierarchy	130
6.4	An Example Concept Taxonomy as Background Knowledge for Non-Taxonomic Relation Extraction	136
6.5	Hierarchical Order on Extracted Non-Taxonomic Relations	138
7.1	TEXT-TO-ONTO Components	153
7.2	TEXT-TO-ONTO Ontology Learning Environment	154
7.3	OntoEdit's View for Lexical Layer Definition	155
7.4	View for Modeling Concepts and Taxonomic Relations	156
7.5	Views for Modeling Non-Taxonomic Relations	157
7.6	View for Modeling Inverse Relations	158
7.7	View for Modeling Disjoint Concepts	159
7.8	ONTOEDIT'S Knowledge Base View	160
7.9	F-Logic Axiom Engineering	161
7.10	View for Querying the SILRI F-Logic Inference Engine	161
7.11	View for Data Selection and Processing	163
7.12	Graphical Interface for Pattern Engineering	164
7.13	Non-Taxonomic Relation Extraction Algorithm View	165
7.14	Result Presentation View	167
7.15	Graph-based Visualization	167
8.1	Levels for Evaluating Ontology Learning	173
8.2	Introduction of Precision and Recall	174
8.3	Example for Computing SC	178
8.4	Two Example Ontologies $\mathcal{O}_1, \mathcal{O}_2$	179
8.5	Example for Computing UC and CM	181
8.6	Two Example Ontologies $\mathcal{O}_1, \mathcal{O}_2$	183
8.7	Measuring Human Modeling Performance	183
8.8	Precision and Recall for Lexical Entry Modeling	185

8.9	Precision and Recall for Concept Hierarchy Modeling	186
8.10	Precision and Recall for Non-Taxonomic Relation Modeling	187
8.11	Measuring Ontology Learning Performance	190
8.12	Precision and Recall for Lexical Entry Extraction	192
8.13	Precision and Recall for Taxonomic Relations Discovery	193
8.14	$\overline{TO}$ of Discovered Taxonomic Relations	194
8.15	Precision and Recall of Non-Taxonomic Relation Discovery	196
9.1	Taxonomy of Related Work	205

List of Tables

3.1	Mapping of $\mathcal{O}$ and $\mathcal{KB}$ to F-Logic	48
5.1	Building an Ontology Wrapper for GermaNet	85
5.2	Example Lexical Entry-Lexical Entry Relation	108
5.3	Example Concept/Lexical Entry-Concept Relation	110
5.4	Example Document-Concept Relation	111
5.5	Example Concept-Transaction Relation	112
5.6	Document Structure Profile	114
6.1	Example Matrix r_{cc}	126
6.2	Examples for Linguistically Related Pairs of Concepts	136
6.3	Examples of Discovered Non-Taxonomic Relations	137
6.4	Example matrix r_{clc}	143
8.1	Basic Statistics – Phase I / Phase II / Phase III	185
8.2	Precision and Recall for Non-Taxonomic Relation Modeling	186
8.3	$\overline{SM}(\mathcal{L}^c_i, \mathcal{L}^c_j), \overline{SM}(\mathcal{L}^R_i, \mathcal{L}^R_j)$ for Phase I-Ontologies.	187
8.4	Typical String Matches	188
8.5	$\overline{TO}(\mathcal{O}_i, \mathcal{O}_j), \overline{RO}(\mathcal{O}_i, \mathcal{O}_j)$ for Phase I-Ontologies.	189
8.6	$\overline{TO}(\mathcal{O}_i, \mathcal{O}_j), \overline{RO}(\mathcal{O}_i, \mathcal{O}_j)$ for Phase II-Ontologies.	189
8.7	$\overline{RO}(\mathcal{O}_i, \mathcal{O}_j)$ for Phase III-Ontologies.	189
8.8	Number of Proposed Lexical Entries	192
8.9	Evaluation Results for Non-Taxonomic Relation Extraction	195
9.1	Example Categorization	216

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, interpreting 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, this work would not have been possible.

I thank my 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 associations 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 multi-strategy 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