Data-Centric Systems and Applications

For further volumes:
http://www.springer.com/series/5258

Zohra Bellahsene • Angela Bonifati
Erhard Rahm
Editors

Schema Matching and Mapping

Springer

Editors
Zohra Bellahsene
LIRMM CNRS/Univ. Montpellier 2
Rue Ada 161
34392 Montpellier
France
bella@lirmm.fr

Angela Bonifati
Consiglio Nazionale delle
Ricerche (CNR)
Via P. Bucci 41/C
87036 Rende
Italy
bonifati@icar.cnr.it

Erhard Rahm
Universität Leipzig
Inst. Informatik
Augustusplatz 10-11
04109 Leipzig
Germany
rahm@informatik.uni-leipzig.de

ISBN 978-3-642-16517-7 e-ISBN 978-3-642-16518-4
DOI 10.1007/978-3-642-16518-4
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2011922131

ACM Computing Classification (1998): H.2, I.2, F.4

Cover design: KuenkelLopka GmbH

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

This book provides an overview about the state-of-the-art solutions and the most recent advances in schema matching and mapping, both recognized as key areas of metadata management. Tasks involving metadata are indeed pervasive in databases and information systems and include schema evolution, schema and ontology integration and matching, XML message mapping, as well as data migration and data exchange. While research on these complex problems has been performed since several decades, we have witnessed significant progress especially in the last decade. In particular, research addressed the metadata problems in a more abstract and generic way rather than focusing on specific applications and data models. A cornerstone of this new line of research is the notion of *schema mappings*, i.e., expressive mappings interrelating schemas (or other metadata models such as ontologies). Furthermore, powerful operators to manipulate schemas and mappings (e.g., matching and merging of schemas or composition of mappings) have been investigated for solving various kinds of metadata-related tasks. Raising the level of abstraction for metadata management was a vision first articulated by Phil Bernstein et al. in *A vision for management of complex models, ACM Sigmod Record 2000.* Since then, many steps have been performed towards the various goals of matching and mapping different kinds of design artifacts (i.e., a relational schema, a web site, or a data mart), thus motivating a flurry of recent research, which we survey in this book. The book consists of ten comprehensive chapters grouped within three parts: large-scale and knowledge-driven schema matching, quality-driven schema mapping and evolution and evaluation and tuning of matching tasks.

The first part deals with schema matching, i.e., the semi-automatic finding of semantic correspondences between elements of two schemas or two ontologies. Schema matching implements a Match operator that is often the first step to determine schema mappings, e.g., for schema evolution, data integration and data exchange. The typically high semantic heterogeneity of the schemas makes schema matching an extremely difficult problem. The separation of Match from other metadata management tasks such as Merge helped to address the match problem better than in the past. Numerous powerful prototypes for schema and ontology matching have been developed in the last decade and automatic match functionality found already its way into commercial products. The four chapters in the first

part cover the achieved state of the art and point out areas where more work is needed, in particular support for large-scale match problems and improved user interaction. Further chapters deal with proposed extensions to enhance the semantic power of match correspondences and to deal with the uncertainty of match decisions.

The second part of the book also consists of four chapters and focuses on schema mappings and their use for schema evolution and schema merging. The first chapter of the second part surveys the existing schema mapping algorithms and the most recent developments towards realizing efficient, optimized and correct schema mapping transformations. Two further chapters deal with the use of schema mappings for schema evolution. One of these introduces the requirements for effective schema evolution support and provides an overview of proposed evolution approaches for diverse kinds of schemas and ontologies. The other evolution-related chapter focuses on the automatic adaptation of mappings after schema changes by presenting two first-class operators on schema mappings, namely composition and inversion. The final chapter surveys the state of the art on mapping-based merging of schemas by discussing the key works in this area and identifying their commonalities and differences.

The third part of the book consists of two chapters devoted to the evaluation and tuning of schema matching and mapping systems. The first of these chapters provides a comprehensive overview of existing evaluation efforts for data transformation tasks, by providing a brand-new perspective under which the various approaches are being/have been evaluated. Such perspective allows the authors to identify the pitfalls of current evaluations and brings them to discuss open problems for future research in this area. The last chapter deals with the complex problem of tuning schema matching tools to optimize their quality and efficiency with a limited amount of configuration effort. An overview of proposed tuning efforts including the use of machine learning techniques is provided.

To the best of our expectations, this book provides:

1. A comprehensive survey of current and past research on schema matching and mapping.
2. An up-to-date source of reference about schema and ontology evolution and schema merging.
3. Scholarly written chapters enabling a learning experience to both experts and non-experts whenever they would like to enhance their knowledge or build it from the scratch; the chapters have been conceived in such a way to be readable individually or altogether by following the book table-of-contents.

As such, we hope that the book proves to be a useful reference to researchers as well as graduate students and advanced professionals. We thank the editors of the DCSA book series, Mike Carey and Stefano Ceri, for their support of our book project and all authors for preparing their chapters and revisions within a few months. Without them, this project would not have been possible. Further thanks go the referees

of the individual chapters for their insightful comments and to Ralf Gerstner from Springer-Verlag for his professional assistance during all the stages of the book production.

September 2010

Zohra Bellasehne
Angela Bonifati
Erhard Rahm

Contents

Part III Evaluating and Tuning of Matching Tasks

Contributors

Zohra Bellahsène LIRMM – CNRS/Université Montpellier II, 161 Rue Ada, 34095 Montpellier Cedex 5, France, bella@lirmm.fr

Angela Bonifati ICAR-CNR, Rende, Italy
bonifati@icar.cnr.it

Xin Luna Dong Data Management Department, AT&T Labs – Research, Bld 103, Rm B281, 180 Park Ave., Florham Park, NJ 07932, USA, lunadong@research.att.com

Fabien Duchateau CWI, Amsterdam, The Netherlands
fabien@cwi.nl

Ronald Fagin IBM Almaden Research Center, Dept. K53/B2, 650 Harry Road, San Jose, CA 95120, USA, fagin@almaden.ibm.com

Sean Falconer Stanford University, Stanford, CA 94305-5479, USA
sean.falconer@stanford.edu

Avigdor Gal Faculty of Industrial Engineering and Management, Technion – Israel Institute of Technology, Technion City, Haifa 32000, Israel, avigal@ie.technion.ac.il

Alon Halevy Google Inc., 1600 Amphitheatre Blvd, Mountain View, CA 94043, USA, halevy@google.com

Michael Hartung Department of Computer Science, University of Leipzig, P.O. Box 100920, 04109 Leipzig, Germany, hartung@informatik.uni-leipzig.de

Phokion Kolaitis IBM Almaden Research Center, 650 Harry Road, San Jose, CA 95120, USA
and
University of California Santa Cruz, 1156 High Street, Santa Cruz, CA 95064, USA, kolaitis@cs.ucsc.edu

Giansalvatore Mecca Dipartimento di Matematica e Informatica, Università della Basilicata, c.da Macchia Romana, 85100 Potenza, Italy, giansalvatore.mecca@unibas.it

Natalya Noy Stanford University, Stanford, CA 94305-5479, USA
and
Medical School Office Building, Room X-215, 251 Campus Drive, Stanford, CA 94305-5479, USA, noy@stanford.edu

Paolo Papotti Dipartimento di Informatica e Automazione, Università Roma Tre, Via della Vasca Navale 79, 00146 Rome, Italy, papotti@dia.uniroma3.it

Lucian Popa 8CC/B1, IBM Almaden Research Center, 650 Harry Road, San Jose, CA 95120, USA, lucian@almaden.ibm.com

Rachel Pottinger Department of Computer Science, University of British Columbia, 201-2366 Main Mall, Vancouver, BC, Canada V6T 1Z4, rap@cs.ubc.ca

Erhard Rahm Department of Computer Science, University of Leipzig, P.O. Box 100920, 04109 Leipzig, Germany, rahm@informatik.uni-leipzig.de

Anish Das Sarma Yahoo! Research, 2-GA 2231, Santa Clara, CA 95051, USA, anish@yahoo-inc.com

Wang-Chiew Tan E3-406, IBM Almaden Research Center, 650 Harry Road, San Jose, CA 95120, USA
and
University of California Santa Cruz, 1156 High Street, Santa Cruz, CA 95064, USA, wctan@cs.ucsc.edu

James F.Terwilliger Microsoft Research, Redmond, WA, USA, James. Terwilliger@microsoft.com

Yannis Velegrakis DISI – University of Trento, Via Sommarive 14, 38123 Trento, Italy, velgias@disi.unitn.eu