Lecture Notes in Computer Science 2917

Edited by G. Goos, J. Hartmanis, and J. van Leeuwen

Springer
Berlin
Heidelberg
New York
Hong Kong
London
Milan
Paris
Tokyo

Elisa Quintarelli

Model-Checking Based Data Retrieval

An Application to Semistructured and Temporal Data

Springer

Series Editors

Gerhard Goos, Karlsruhe University, Germany
Juris Hartmanis, Cornell University, NY, USA
Jan van Leeuwen, Utrecht University, The Netherlands

Author

Elisa Quintarelli
Politecnico di Milano
Dip. di Elettronica e Informazione
Piazza Leonardo da Vinci 32, 20133 Milano, Italy
E-mail: quintare@elet.polimi.it

Cataloging-in-Publication Data applied for

A catalog record for this book is available from the Library of Congress.

Bibliographic information published by Die Deutsche Bibliothek
Die Deutsche Bibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data is available in the Internet at <http://dnb.ddb.de>.

CR Subject Classification (1998): H.2, H.3, H.4

ISSN 0302-9743
ISBN 3-540-20971-9 Springer-Verlag Berlin Heidelberg New York

Typesetting: Camera-ready by author, data conversion by Boller Mediendesign
Printed on acid-free paper SPIN: 10977132 06/3142 5 4 3 2 1 0

To Alessia

Foreword

This thesis deals with the problems of characterizing the semantics of and assuring efficient execution for database query languages, where the database contains semistructured and time-varying information. This area of technology is of much interest and significance for databases and knowledge bases; it also presents many challenging research problems deserving an in-depth investigation. Thus, the topic of Elisa Quintarelli's dissertation is well chosen and totally appropriate to the current research trends.

In her thesis, Elisa addresses a number of related problems. However, her work and contributions concentrate on two main problems. The first is the definition of an effective graph-based approach to the formalization of query languages for semistructured and temporal information. In her approach, query execution is viewed as the process of matching the query graph with the database instance graph; therefore, query execution reduces to searching the database for subgraphs that are similar to the given query graph. The search for such matches can be supported through the computational process of bisimulation. This approach is used to define the semantics of several languages, including graphical languages, such as G-Log and GraphLog, semistructured information languages, such as Lorel, and temporal languages, such as TSS-QL. Both graph-based approaches and bisimulation had been used by previous authors for defining query languages and their semantics; however, this work goes well beyond previous approaches by integrating and refining these techniques into a flexible and powerful paradigm that Elisa demonstrates to be effective on a spectrum of languages and a suite of alternative semantics.

The second research challenge tackled by Elisa in her thesis is that of efficient implementation. This is a nontrivial problem since bisimulation can, in the worst case, incur an exponential time complexity. Her original solution to this difficult problem consists of modeling graphical queries as formulas of modal logic and interpreting database instance graphs as Kripke transition systems. In this way, the problem of solving graphical queries is reduced to the model-checking problem for which efficient decision algorithms exist. In particular, the thesis focuses on CTL formulae for which efficient model checkers are available; this allows Elisa to demonstrate the efficiency of her proposed approach with experimental results.

In conclusion, the thesis represents an interesting piece of research, characterized by novel contributions and an in-depth expertise on several topics. Indeed, the author brings together ideas and techniques from different areas, and displays a solid expertise in computing, in general, and information systems, in particular. In my role as her advisor I had the privilege to see Elisa's growth as a researcher; overall, I am still impressed with the depth and breadth of her work, the originality of her results, and her research maturity. Therefore, I am very happy to see that her Ph.D. thesis has been chosen to be published as a book in these Lecture Notes in Computer Science.

Milan, 10/10/2003 *Letizia Tanca*

Preface

This book contains the research covered by my Ph.D. Thesis, which was developed at the Dipartimento di Elettronica of the Politecnico di Milano, in collaboration with Prof. Agostino Dovier at the Dipartimento di Informatica of the Università di Verona.

The main topic of the book is the study of appropriate, flexible and efficient search techniques for querying semistructured and WWW data, also taking into account the time dimension.

This work was motivated by the fact that in recent years a lot of attention has been given by the database research community to the introduction of methods for representing and querying *semistructured data*. Roughly speaking, this term is used for data that have no absolute schema fixed in advance, and whose structure may be irregular or incomplete.

A common example in which semistructured data arise is when data are stored in sources that do not impose a rigid structure, such as the World Wide Web, or when they are extracted from multiple heterogeneous sources. It is evident that an increasing amount of semistructured data is becoming available to users and, thus, there is a need for Web-enabled applications to access, query and process heterogeneous or semistructured information, flexibly dealing with variations in their structure.

This work proposes a formalization to provide suitable graph-based semantics to languages for semistructured data, supporting both data structure variability and topological similarities between queries and document structures. A suite of semantics based on the notion of bisimulation is introduced both at the concrete level (instances) and at the abstract level (schemata) for a graph-based query language, but the results are general enough to be adapted to other existing proposals as well. Moreover, complexity results on the matching techniques, which are required to find a sort of similarity between a graphical query and a semistructured database, have stimulated our interest in investigating alternative approaches to solve such graphical queries on databases. The main idea here is to solve the data retrieval problem for semistructured data by using techniques and algorithms coming from the model-checking research field: experimental results are presented in this work to confirm the possibility of effectively applying the proposed method based on model-checking algorithms.

The book is structured in the following way: Chap. 1 describes the current scenario, the motivations and the contributions of this work.

Chapter 2 sets the formal content, by presenting G-Log, a graph-based query language showing a three-level semantics based on the notion of bisimulation; the relationships between instances and schemata are investigated by using the theory of abstract interpretation, which provides a systematic approach to guarantee the correctness of operating on schemata with respect to the corresponding concrete computations on instances. We chose G-Log because it is a general graphical language that combines the expressive power of logic, the modeling power of objects, and the representation power of graphs for instances, schemata and logical inferences (i.e., queries and rules). The chapter also describes the main features of two well-known query languages, namely Graph-Log and UnQL, in order to show the applicability and generality of the proposed study to other SQL-like and graphical languages for semistructured data.

Chapter 3 describes the novel idea of this work. We propose an approach to associate a *modal* logic formula to a graphical query, and to interpret database instance graphs as *Kripke Transition Systems* (*KTS*). In this way, the problem of finding subgraphs of the database instance graph that match the query can be performed by using *model-checking* algorithms. In particular, we identify a family of graph-based queries that represent *CTL* formulae. This is very natural and, as an immediate consequence, an effective procedure for efficiently querying semistructured databases can be directly implemented on a model-checker and the query retrieval activity can be performed in polynomial time. In Chap. 3 we focus on a graphical query language very similar to G-Log: we consider only the main constructs of this language that have a natural translation into temporal logic, but we emphasize the possibility of expressing universal conditions in a very compact way. In this chapter we also show some experimental results that confirm our proposal.

Starting with the potential of the *CTL* language for time-based representation, Chap. 4 extends standard techniques for modeling and accessing static information in order to model and retrieve temporal aspects of semistructured data, such as, for example, evolution on simple values contained in semistructured documents. In particular we present a generic graphical data model suitable for representing both static and dynamic aspects of semistructured data, and we introduce a very simple SQL-like query language to compose temporal inferences. In this chapter we do not extend G-Log because it is more natural to express temporal conditions with SQL properties; however, we informally propose the possibility of encoding such properties into graphical queries by adapting the model originally introduced for G-Log in a very natural way. We show also the fragment of this SQL-like language that can be translated into the logic *CTL*, and thus we propose applying model-checking algorithms to solve temporal queries as well. In the same chapter we concentrate our attention not only on the classical notion of the time concept,

as it is deeply known in the database field, but we further investigate the possibility of considering more than one dynamic aspect of semistructured data. In particular, we represent with the same data model valid time and interaction time. The first notion is used to consider the validity of facts in the represented reality, whereas the second one is related to user browsing history while navigating Web sites.

Finally, in Chap. 5 we compare this work with others known in the literature, while Chap. 6 summarizes the main results of the thesis and proposes further developments of our piece of research.

Acknowledgments

At the end of this experience it seems only right and proper to give special mention to the people who accompanied me on this long trip.

I would like to take this opportunity to express my deepest thanks to my advisor, Letizia Tanca, for her immense help and encouragement over the last three years. Thanks for always being ready with advice on any possible problem (from scientific questions to very personal problems) and for your efforts to improve this book.

A special thanks to Agostino Dovier for all the things he taught me: how to think and do research, how to write a paper, and how to deal with a formal proof. Thanks for the many corrections to the things we wrote together.

I would like to thank my examiner, Prof. Carlo Zaniolo, for his comments and suggestions on the manuscript.

I am also indebted to Ernesto Damiani for his contribution to some aspects of this work, and to Fabio Grandi for the stimulating discussion on temporal databases during Time 2001. I thank also Carlo Combi for the help he has given me on time-related topics.

I would like to thank Barbara, for being my best friend, for the times she shared with me, for making me laugh on the many occasions she knows very well, and for having convinced me to make some decisions . . . Thanks Barbara for all the experiences we shared together during this Ph.D., I will always treasure them.

I thank Nico for challenging me with the first steps in the model-checking field.

I would like to say "thank you" to the Ph.D. students and friends of the DEI department, who have made my life on the first floor most pleasant. Among them, I especially want to mention Vincenzo, Giovanni, Mattia, Matteo, Marco, and Fabio for their enjoyable company, and Sara for her constant support and friendship.

I cannot forget Angela, for her efforts to improve my English, and some special friends: Nadia and Francesca.

I would like to express my gratitude to Rosalba for being a true friend to me.

I would like to give a special thanks to my family: to my parents Federico and Paola for raising me, and for encouraging me to go on with my studies.

This thesis is also for my sisters Giulia and Marta; their constant love and support was very important to me.

Last, but definitely not least, I would like to thank Massimo, for sharing all my achievements with me, for his precious advice in difficult moments, and especially for keeping me connected to the real world; thanks for being with me at all times. A very special thanks to my lovely baby Alessia who makes every day of my life unpredictable.

Milan, 10/10/2003 *Elisa Quintarelli*

Contents