Lecture Notes in Computer Science 2261

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

Springer
Berlin
Heidelberg
New York
Barcelona
Hong Kong
London
Milan
Paris
Tokyo

Felix Naumann

Quality-Driven Query Answering for Integrated Information Systems

Springer

Series Editors

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

Author

Felix Naumann
IBM Almaden Research Center
650 Harry Road, San Jose, CA 95123, USA
E-mail: felix@almaden.ibm.com

Dissertation der Humboldt Universität zu Berlin
Tag der mündlichen Prüfung: 19. Dezember 2000

Referent: Prof. Dr. Johann-Christoph Freytag, Humboldt Universität zu Berlin
Referent: Prof. Myra Spiliopoulou, Handelshochschule Leipzig
Referent: Prof. Gio Wiederhold, Stanford University

Cataloging-in-Publication Data applied for

Die Deutsche Bibliothek - CIP-Einheitsaufnahme

Naumann, Felix:
Quality driven query answering for integrated information systems /
Felix Naumann. - Berlin ; Heidelberg ; New York ; Barcelona ; Hong Kong ;
London ; Milan ; Paris ; Tokyo : Springer, 2002
(Lecture notes in computer science ; 2261)
ISBN 3-540-43349-X

CR Subject Classification (1998): H.3, H.2, H.4, I.2.11, C.2.4, F.2.2

ISSN 0302-9743
ISBN 3-540-43349-X Springer-Verlag Berlin Heidelberg New York

Springer-Verlag Berlin Heidelberg New York
a member of BertelsmannSpringer Science+Business Media GmbH

http://www.springer.de

Printed in Germany

Typesetting: Camera-ready by author, data conversion by Boller Mediendesign
Printed on acid-free paper SPIN: 10846092 06/3142 5 4 3 2 1 0

Foreword

The Internet and the World Wide Web (WWW) are becoming more and more important in our highly interconnected world as more and more data and information is made available for online access. Many individuals and governmental, commercial, cultural, and scientific organizations increasingly depend on information sources that can be accessed and queried over the Web. For example, accessing flight schedules or retrieving stock information has become common practice in today's world. When accessing this data, many people assume that the information accessed is accurate and that the data source can be accessed reliably.

These two examples clearly demonstrate that not only the information content is important, the **information about the quality of the data** becomes an even more crucial and critical aspect for individuals and organizations when they make plans or take decisions based on the results of their queries. More precisely, having access to information of known quality becomes critical for the well-being and indeed for the functioning of modern industrialized societies.

Surprisingly, despite the urgent need for clear concepts and techniques to judge and value quality and for technology to use such (meta) information, very few scientific results are known and available. Few approaches are known to use quality measures for accessing and querying information over the Web. Only a limited number of products on the IT market address this burning problem.

With this book Dr. Felix Naumann is one of the first to address the topic of querying data with quality in a systematic and comprehensive way from a database point of view. His two-step approach reflects his clear understanding of the problem as well as the solutions required for "real-world" settings. As a basis he first describes the various "properties" of information quality (IQ) by more specific – and technically sound – quality measures before introducing ranking algorithms to select Web sources for access. In the second part of his book, Dr. Naumann focuses on "quality-driven" query answering, in particular on query planning using the different quality criteria. Again, the solutions presented reflect Dr. Naumann's desire not to come up with *any* solution, but rather to design algorithms that could be used in "real world" systems, a goal he greatly achieves. His particular focus on completeness of data, a very important aspect for "real-world" scenarios, together with the designed algorithms, is another highlight of this book. The careful reader will notice – despite the many technical details – that

Dr. Naumann's in-depth treatment of completeness provides the insight into the problem necessary for such an important topic.

In summary, the approach and systematic treatment of information quality taken in this book and the way Dr. Naumann describes problems and solutions makes this book valuable for both researchers and practitioners who are interested in gaining a better understanding of the issues and solutions available in the context of information quality. The in-depth presentation of the algorithms and techniques is enlightening to students and a valuable resource for computer scientists as well as for business people. I predict that in the years ahead this book will provide the "road map" for others in this area both in research and development.

November 2001 Johann-Christoph Freytag

Preface

Research and business is currently moving from centralized databases towards information systems integrating distributed, autonomous data sources. With it, research focus has shifted from traditional query optimization to the field of query planning. Query planning is the problem of finding query execution plans across distributed, heterogeneous, overlapping, and autonomous data sources. We argue that for such data sources the main discriminator for different query execution strategies is no longer response time, as it is for database queries, but – more generally – the *information quality* (IQ) of the result. This thesis investigates the usage of IQ-criteria to improve the answering of user queries against integrated information systems. We discuss what kind of IQ-metadata is necessary, how it can be acquired, and – most importantly – how it can be used to improve the quality of query results and the performance of query planning algorithms. A simple application for these research issues is a meta-search engine that uses existing search engines as its distributed data sources. Other examples include stock information systems, travel guides, and distributed molecular biology databases.

The thesis has three main parts. Part I lays the foundation for the problem of querying Web data sources and shows why IQ-reasoning is helpful. We describe the mediator–wrapper architecture and show how to describe sources and user queries using the concept of the universal relation. Several application examples serve as rationale throughout the thesis.

Part II introduces our model of information quality. We present a comprehensive set of IQ-criteria together with score assessment methods. Each data source is rated by a set of IQ-criteria, such as completeness, understandability, or accuracy. To compare data sources and query plans qualitatively using multiple criteria, we present appropriate ranking methods, which aggregate IQ-criterion scores to an overall quality value.

Part III puts information quality to work by combining query planning with IQ-reasoning. We revise the conventional query planning goal of finding all plans for a query: The new goal is to find the best N plans and use a quality model to quantify the term 'best'. We present two algorithms to solve this problem. The first acts as an add-on to any given query planning algorithm, the second explicitly integrates IQ-reasoning into the planning process, thereby speeding up query planning itself. Next, we part from the conven-

tional query planning paradigm of finding different plans for a query, each with a different result. The usage of new outerjoin-type merge operators to combine sources enables a reduction of the paradigm to finding a single, best plan. We concentrate on the completeness criterion describing the amount of data returned by a plan and present two families of optimization algorithms for different real world situations. All algorithms are evaluated using a simulation testbed.

The main contribution of the thesis is the comprehensive integration of information quality reasoning and query planning. Research has recognized the importance of quality reasoning, but, to the best of our knowledge, IQ-reasoning for query planning has not been adequately addressed before.

Acknowledgments

I thank my advisors Prof. Johann Christoph Freytag, Prof. Myra Spiliopoulou, and Prof. Hans-Joachim Lenz for their constant support, frequent meetings, and useful advice. Prof. Freytag offered a great research environment that was a pleasure to work in and turned three years of research into rewarding results and an enjoyable time. Myra introduced me to database research, had the first inkling that quality-reasoning is a promising research topic, and guided my research in many meetings. Prof. Lenz helped find different perspectives on my work and never hesitated to plunge into the depths of my research. Also, I thank Prof. Gio Wiederhold for an extensive and valuable discussion and for reviewing this thesis.

The Graduate School for Distributed Information Systems supported this thesis financially and – more importantly – through regular discussions and evaluation by its professors and students. Among the many students of the graduate school to whom I am grateful for three wonderful years, I point out Ulf Leser and André Bergholz. Due to Ulf, my research was particularly enjoyable and successful, through much help, many discussions, and having a great time together. André was an important constant for my research and university life, leading the path I was to follow.

Working with the dbis team at the Humboldt University has been good fun. Ulrike Sholz and Heinz Werner were especially patient and helpful. Mike Stillger gave superb guidance for an inexperienced colleague – thanks Mike. I thank Ramana Yerneni and Prof. Hector Garcia-Molina for an enlightening stay at Stanford University, Claudia Rolker who I have yet to meet, and finally Julia Böttcher and Daniel Tonn for our successful projects.

Last but not least, Anneke, my parents, and my sister supported me and put up with me all along.

This research was supported by the German Research Society, Berlin-Brandenburg Graduate School in Distributed Information Systems (DFG grant no. GRK 316).

August 2001 *Felix Naumann*

Contents