Database Repairing and Consistent Query Answering

Synthesis Lectures on Data Management

Editor
M. Tamer Özsu, ***University of Waterloo***

Synthesis Lectures on Data Management is edited by Tamer Özsu of the University of Waterloo. The series will publish 50- to 125-page publications on topics pertaining to data management. The scope will largely follow the purview of premier information and computer science conferences, such as ACM SIGMOD, VLDB, ICDE, PODS, ICDT, and ACM KDD. Potential topics include, but not are limited to: query languages, database system architectures, transaction management, data warehousing, XML and databases, data stream systems, wide scale data distribution, multimedia data management, data mining, and related subjects.

Database Repairing and Consistent Query Answering
Leopoldo Bertossi
2011

Managing Event Information: Modeling, Retrieval, and Applications
Amarnath Gupta and Ramesh Jain
2011

Fundamentals of Physical Design and Query Compilation
David Toman and Grant Weddell
2011

Methods for Mining and Summarizing Text Conversations
Giuseppe Carenini, Gabriel Murray, and Raymond Ng
2011

Probabilistic Databases
Dan Suciu, Dan Olteanu, Christopher Ré, and Christoph Koch
2011

Peer-to-Peer Data Management
Karl Aberer
2011

Probabilistic Ranking Techniques in Relational Databases
Ihab F. Ilyas and Mohamed A. Soliman
2011

Uncertain Schema Matching
Avigdor Gal
2011

Fundamentals of Object Databases: Object-Oriented and Object-Relational Design
Suzanne W. Dietrich and Susan D. Urban
2010

Advanced Metasearch Engine Technology
Weiyi Meng and Clement T. Yu
2010

Web Page Recommendation Models: Theory and Algorithms
Sule Gündüz-Ögüdücü
2010

Multidimensional Databases and Data Warehousing
Christian S. Jensen, Torben Bach Pedersen, and Christian Thomsen
2010

Database Replication
Bettina Kemme, Ricardo Jimenez Peris, and Marta Patino-Martinez
2010

Relational and XML Data Exchange
Marcelo Arenas, Pablo Barcelo, Leonid Libkin, and Filip Murlak
2010

User-Centered Data Management
Tiziana Catarci, Alan Dix, Stephen Kimani, and Giuseppe Santucci
2010

Data Stream Management
Lukasz Golab and M. Tamer Özsu
2010

Access Control in Data Management Systems
Elena Ferrari
2010

An Introduction to Duplicate Detection
Felix Naumann and Melanie Herschel
2010

Privacy-Preserving Data Publishing: An Overview
Raymond Chi-Wing Wong and Ada Wai-Chee Fu
2010

Keyword Search in Databases
Jeffrey Xu Yu, Lu Qin, and Lijun Chang
2009

Database Repairing and Consistent Query Answering

Leopoldo Bertossi

ISBN: 978-3-031-00755-2 paperback
ISBN: 978-3-031-01883-1 ebook

DOI 10.1007/978-3-031-01883-1

A Publication in the Springer series
SYNTHESIS LECTURES ON DATA MANAGEMENT

Lecture #20
Series Editor: M. Tamer Özsu, *University of Waterloo*
Series ISSN
Synthesis Lectures on Data Management
Print 2153-5418 Electronic 2153-5426

Database Repairing and Consistent Query Answering

Leopoldo Bertossi
Carleton University, Ottawa, Canada

SYNTHESIS LECTURES ON DATA MANAGEMENT #20

ABSTRACT

Integrity constraints are semantic conditions that a database should satisfy in order to be an appropriate model of external reality. In practice, and for many reasons, a database may not satisfy those integrity constraints, and for that reason it is said to be *inconsistent*. However, and most likely a large portion of the database is still semantically correct, in a sense that has to be made precise. After having provided a formal characterization of consistent data in an inconsistent database, the natural problem emerges of extracting that semantically correct data, as query answers.

The consistent data in an inconsistent database is usually characterized as the data that persists across all the database instances that are consistent and minimally differ from the inconsistent instance. Those are the so-called *repairs* of the database. In particular, the *consistent answers* to a query posed to the inconsistent database are those answers that can be simultaneously obtained from all the database repairs.

As expected, the notion of repair requires an adequate notion of distance that allows for the comparison of databases with respect to how much they differ from the inconsistent instance. On this basis, the minimality condition on repairs can be properly formulated.

In this monograph we present and discuss these fundamental concepts, different repair semantics, algorithms for computing consistent answers to queries, and also complexity-theoretic results related to the computation of repairs and doing consistent query answering.

KEYWORDS

integrity constraints, inconsistent databases, database repairs, consistent query answering, data cleaning

To Prof. Jöerg Flum,
for his guidance, support, an example of scholarship,
relentless research activity, and humanity

Contents

Preface

A common assumption in data management is that databases can be kept *consistent*, that is, satisfying certain desirable integrity constraints (ICs). This is usually achieved by means of built-in support provided by database management systems. They allow for the *maintenance* of limited classes of ICs that can be declared together with the database schema. Another possibility is the use of *triggers* or *active rules* that are created by the user and stored in the database. They react to updates of the database by notifying a violation of an IC, rejecting a violating update, or compensating the update with additional updates that restore consistency. Another common alternative consists of keeping the ICs satisfied through the application programs that access and modify the database, i.e., from the transactional side.

However, under different circumstances and for several reasons, databases may be or may become inconsistent. For example, ICs that are expensive to check and maintain, enforcement or simple consideration of new or user ICs, imposition of a new semantics on legacy data, the creation of a repository of integrated data, etc. Confronted to the possible or potential inconsistency of a database, we may decide to live with this inconsistency, but trying to access, retrieve and use the portion of data that is still consistent with respect to the ICs under consideration.

Consistent query answering (CQA) [Arenas et al., 1999] emerged from this attitude towards inconsistency and the need to do semantically correct data management, in particular, query answering, in the presence of inconsistency. This required a precise, formal characterization of the consistent data in a possibly inconsistent database, and also the development of computational mechanisms for retrieving the consistent data, e.g., at query answering time.

The characterization of consistent data, as first proposed by Arenas et al. [1999], appeals to the auxiliary notion of *database repair*. This is a new database instance that is consistent with respect to the ICs, and *minimally differs* from the inconsistent database at hand. Consistent data is invariant under the class of possible repairs. Since their official inception, CQA has received much attention from the research community in data management. The main problems mentioned above, i.e., characterization of consistent data and the development of efficient algorithms, have been largely explored. The former under different notions of repair (and distance between instances), and the latter, considering all kinds of combinations of classes of ICs and queries, including complexity-theoretic issues.

In this monograph we introduce the motivation, the main concepts and techniques, and also the main research problems that appear behind and around database repairs and CQA. Much research has been produced and published in the last 12 years. It would be impossible to give a detailed account of it in a rather short monograph like this. As a consequence, the treatment of most of the topics and research results is kept rather intuitive and superficial, but hopefully still precise enough. We have preferred illustrative and representative examples to full proofs of theorems. However, we have provided abundant references to the publications where those results can be found in full detail, and much more. Some surveys of the area have been published before [Bertossi, 2006, Bertossi and Chomicki, 2003, Chomicki, 2007].

This monograph concentrates on CQA and database repairs in/for single relational databases. As a consequence, some topics in CQA and repairs for other data models have been omitted. Some of them are mentioned below.

Consistent query answering and repairs for *XML databases* have been considered by Flesca et al. [2005a,b] and Staworko and Chomicki [2006]. The same problems in *multidimensional databases* (MDDBs), but with semantic constraints like *homogeneity* and *strictness*, have been considered by Bertossi et al. [2009], Bravo et al. [2010] and Ariyan and Bertossi [2011], on the basis of the Hurtado-Mendelzon model for MDDBs [Hurtado and Mendelzon, 2002]. And for *spatial databases*, by Rodriguez et al. [2008, 2011].

Consistent query answering and database repairs has been applied in *virtual data integration systems* that are subject to global integrity constraints [Bertossi and Bravo, 2004b, Bravo and Bertossi, 2003, 2005]. They have also played an important role in *peer data exchange systems* that exchange data at query answering time when certain data exchange constraints between peers are violated. In consequence, inconsistency is the driving force behind data movement between peers [Bertossi and Bravo, 2004a, 2007, 2008].

We are not presenting here research on *probabilistic* representation of repairs or repairs in *probabilistic databases* [Andritsos et al., 2006, Lian et al., 2010] (except for general remarks in Section 6.2.1).

Leopoldo Bertossi
August 2011

Acknowledgments

I am indebted to many people who have made writing this monograph possible. Special thanks go to my several coauthors of papers on consistent query answering, database repairs, and data cleaning. It has been a stimulating, enriching and pleasant experience working with them.

I am especially grateful to Jef Wijsen, Ihab Ilyas, and Solmaz Kolahi for their help with contributed material to this monograph. Also to Filippo Furfaro and Franceso Parisi for personal conversations and some useful material on their research. Valuable and detailed comments and suggestions by Jan Chomicki to a first version of this monograph are very much appreciated. Of course, in the end I am the only one responsible for any errors introduced, and also for the choice of subjects and the way they are presented.

I am grateful to Tamer Özsu for the opportunity to write this monograph, and his suggestions for improving it. Interacting with Tamer is always a pleasure.

My warmest thanks go to my wife, Jennifer, and my daughter, Paloma, for all the support, patience, and love I have received from them during all the time, several years by now, that I have spent doing research in the area.

Leopoldo Bertossi
August 2011