

Answering Queries Using Views

Second Edition

Synthesis Lectures on Data Management

Editor

H.V. Jagadish, *University of Michigan*

Founding Editor

M. Tamer Özsu, *University of Waterloo*

Synthesis Lectures on Data Management is edited by H.V. Jagadish of the University of Michigan. The series publishes 80–150 page publications on topics pertaining to data management. Topics include 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.

Answering Queries Using Views, Second Edition

Foto Afrati and Rada Chirkova
2019

Data Exploration Using Example-Based Methods

Matteo Lissandrini, Davide Mottin, Themis Palpanas, and Yannis Velegrakis
2018

Data Profiling

Ziawasch Abedjan, Lukasz Golab, Felix Naumann, and Thorsten Papenbrock
2018

Querying Graphs

Angela Bonifati, George Fletcher, Hannes Voigt, and Nikolay Yakovets
2018

Query Processing over Incomplete Databases

Yunjun Gao and Xiaoye Miao
2018

Natural Language Data Management and Interfaces

Yun Yao Li and Davood Rafiei
2018

Human Interaction with Graphs: A Visual Querying Perspective

Sourav S. Bhowmick, Byron Choi, and Chengkai Li
2018

On Uncertain Graphs

Arijit Khan, Yuan Ye, and Lei Chen
2018

Answering Queries Using Views

Foto Afrati and Rada Chirkova
2017

Databases on Modern Hardware: How to Stop Underutilization and Love Multicores

Anatasia Ailamaki, Erieta Liarou, Pinar Tözün, Danica Porobic, and Iraklis Psaroudakis
2017

Instant Recovery with Write-Ahead Logging: Page Repair, System Restart, Media Restore, and System Failover, Second Edition

Goetz Graefe, Wey Guy, and Caetano Sauer
2016

Generating Plans from Proofs: The Interpolation-based Approach to Query Reformulation

Michael Benedikt, Julien Leblay, Balder ten Cate, and Efthymia Tsamoura
2016

Veracity of Data: From Truth Discovery Computation Algorithms to Models of Misinformation Dynamics

Laure Berti-Équille and Javier Borge-Holthoefer
2015

Datalog and Logic Databases

Sergio Greco and Cristina Molinaro
2015

Big Data Integration

Xin Luna Dong and Divesh Srivastava
2015

Instant Recovery with Write-Ahead Logging: Page Repair, System Restart, and Media Restore

Goetz Graefe, Wey Guy, and Caetano Sauer
2014

Similarity Joins in Relational Database Systems

Nikolaus Augsten and Michael H. Böhlen
2013

Information and Influence Propagation in Social Networks

Wei Chen, Laks V.S. Lakshmanan, and Carlos Castillo
2013

Data Cleaning: A Practical Perspective

Venkatesh Ganti and Anish Das Sarma
2013

Data Processing on FPGAs

Jens Teubner and Louis Woods
2013

Perspectives on Business Intelligence

Raymond T. Ng, Patricia C. Arocena, Denilson Barbosa, Giuseppe Carenini, Luiz Gomes, Jr.,
Stephan Jou, Rock Anthony Leung, Evangelos Milios, Renée J. Miller, John Mylopoulos, Rachel
A. Pottinger, Frank Tompa, and Eric Yu
2013

Semantics Empowered Web 3.0: Managing Enterprise, Social, Sensor, and Cloud-based Data and Services for Advanced Applications

Amit Sheth and Krishnaprasad Thirunarayan
2012

Data Management in the Cloud: Challenges and Opportunities

Divyakant Agrawal, Sudipto Das, and Amr El Abbadi
2012

Query Processing over Uncertain Databases

Lei Chen and Xiang Lian
2012

Foundations of Data Quality Management

Wenfei Fan and Floris Geerts
2012

Incomplete Data and Data Dependencies in Relational Databases

Sergio Greco, Cristian Molinaro, and Francesca Spezzano
2012

Business Processes: A Database Perspective

Daniel Deutch and Tova Milo
2012

Data Protection from Insider Threats

Elisa Bertino
2012

Deep Web Query Interface Understanding and Integration

Eduard C. Dragut, Weiyi Meng, and Clement T. Yu

2012

P2P Techniques for Decentralized Applications

Esther Pacitti, Reza Akbarinia, and Manal El-Dick

2012

Query Answer Authentication

HweeHwa Pang and Kian-Lee Tan

2012

Declarative Networking

Boon Thau Loo and Wenchao Zhou

2012

Full-Text (Substring) Indexes in External Memory

Marina Barsky, Ulrike Stege, and Alex Thomo

2011

Spatial Data Management

Nikos Mamoulis

2011

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

© Springer Nature Switzerland AG 2022

Reprint of original edition © Morgan & Claypool 2019

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—electronic, mechanical, photocopy, recording, or any other except for brief quotations in printed reviews, without the prior permission of the publisher.

Answering Queries Using Views, Second Edition

Foto Afrati and Rada Chirkova

ISBN: 978-3-031-00743-9 paperback

ISBN: 978-3-031-01871-8 ebook

ISBN: 978-3-031-00098-0 hardcover

DOI 10.1007/978-3-031-01871-8

A Publication in the Springer series

SYNTHESIS LECTURES ON DATA MANAGEMENT

Lecture #54

Series Editor: H.V. Jagadish, *University of Michigan*

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

Series ISSN

Print 2153-5418 Electronic 2153-5426

Answering Queries Using Views

Second Edition

Foto Afrati

National Technical University of Athens

Rada Chirkova

North Carolina State University

SYNTHESIS LECTURES ON DATA MANAGEMENT #54

ABSTRACT

The topic of using views to answer queries has been popular for a few decades now, as it cuts across domains such as query optimization, information integration, data warehousing, website design and, recently, database-as-a-service and data placement in cloud systems.

This book assembles foundational work on answering queries using views in a self-contained manner, with an effort to choose material that constitutes the backbone of the research. It presents efficient algorithms and covers the following problems: query containment; rewriting queries using views in various logical languages; equivalent rewritings and maximally contained rewritings; and computing certain answers in the data-integration and data-exchange settings. Query languages that are considered are fragments of SQL, in particular select-project-join queries, also called conjunctive queries (with or without arithmetic comparisons or negation), and aggregate SQL queries.

This second edition includes two new chapters that refer to tree-like data and respective query languages. Chapter 8 presents the data model for XML documents and the XPath query language, and Chapter 9 provides a theoretical presentation of tree-like data model and query language where the tuples of a relation share a tree-structured schema for that relation and the query language is a dialect of SQL with evaluation techniques appropriately modified to fit the richer schema.

KEYWORDS

conjunctive queries, aggregate queries, arithmetic comparisons, negation, acyclic queries, query containment, query equivalence, equivalent query rewritings, maximally contained query rewritings, tuple-generating dependencies, equality-generating dependencies, the chase algorithm query containment and rewriting under dependencies, data exchange, determinacy, XPath, tree data

Contents

	Preface to the First Edition	xvii
	Preface to the Second Edition	xix
	Acknowledgments	xxi
1	Queries and Views	1
1.1	Views	1
1.2	Using Views in Database Systems	2
1.2.1	Views and Use Cases	2
1.2.2	Views and Technical Issues	3
1.3	Answering Queries Using Views	4
1.3.1	Producing All the Answers—Equivalent Rewritings	5
1.3.2	Producing a Subset of the Answers—Contained Rewritings	7
1.3.3	Data Exchange	8
1.4	Relational Databases and Queries	9
1.4.1	Conjunctive Queries	10
1.4.2	Conjunctive Queries with Arithmetic Comparisons (CQAC)	13
1.4.3	Conjunctive Queries with Negation (CQN)	14
1.4.4	Unions of Conjunctive Queries	14
1.4.5	Conjunctive Queries with Aggregation (CQA)	15
1.5	The Structure of the Book	17
1.6	Exercises	19
2	Query Containment and Equivalence	21
2.1	CQ Query Containment	21
2.1.1	Containment Mapping and Homomorphisms: Canonical Database ..	22
2.1.2	Query Evaluation vs. Query Containment	24
2.2	CQAC Query Containment	26
2.2.1	Multiple Containment Mappings and Normalization: Set of Canonical Databases	27
2.2.2	When Normalization is not Needed	32

2.2.3	When Single Mapping Suffices—the Homomorphism Property: AC-canonical Databases	33
2.3	CQN Query Containment	37
2.3.1	Set of Canonical Databases	37
2.4	CQA Query Containment and Equivalence	40
2.4.1	MAX Queries	40
2.4.2	SUM Queries	43
2.4.3	More General Aggregate Operators	44
2.5	Acyclic CQs	45
2.5.1	Definition of Acyclic Queries and Special Cases	45
2.5.2	Efficient Algorithm for Checking Query Containment	48
2.5.3	Efficient Algorithm for Query Evaluation	51
2.6	Query Equivalence	53
2.7	Containment and Equivalence for Unions of Queries	54
2.8	Exercises	56
3	Finding Equivalent Rewritings	61
3.1	Preliminaries	61
3.1.1	Expansion of a Rewriting	62
3.1.2	Contained Rewritings	64
3.2	CQ Queries and Views	65
3.2.1	Canonical Rewritings and the Naive Algorithm	65
3.2.2	Properties of the Expansion	68
3.2.3	Algorithm CoreCover	71
3.3	Acyclic CQ Views	73
3.4	CQAC Queries and Views	73
3.4.1	Algorithms for Finding Equivalent Rewritings for Queries and Views with ACs	74
3.4.2	When the Homomorphism Property Holds	75
3.5	Rewriting CQN Queries Using CQN Views	77
3.6	CQA Queries	80
3.6.1	Unfoldings of Rewritings	81
3.6.2	Equivalence of Unfoldings and Rewritings	81
3.6.3	Constructing Central Rewritings	84
3.6.4	Non-central Rewritings	86
3.7	Exercises	87

4	Maximally Contained Rewritings (MCRs)	89
4.1	Preliminaries	90
4.1.1	CQAC Queries and Views	92
4.2	Finding MCRs for CQ Queries and Views: The MS Algorithm	95
4.2.1	Finding MCDs	97
4.2.2	Property of MCDs	99
4.2.3	Combining MCDs—Most Relaxed Rewriting	100
4.3	CQACs, the Homomorphism Property, Extending Algorithm MS	102
4.3.1	Exportable and Distinguishable Variables of an Expansion	104
4.3.2	AC-MCDs	106
4.3.3	Combining AC-MCDs and MCDs, and Adding ACs	107
4.4	Datalog	110
4.4.1	Definition of the Datalog Language	111
4.4.2	Views that are Unions of CQs	113
4.4.3	Rewriting Datalog Queries—the Inverse-rule Algorithm	114
4.4.4	CQAC-SI Queries and Views	121
4.5	Exercises	122
5	Answering Queries in Presence of Dependencies	125
5.1	Preliminaries	125
5.1.1	The Chase	126
5.1.2	Weakly Acyclic Sets of Tgds	127
5.1.3	Property of the Chase	129
5.2	Query Containment Under Dependencies	131
5.3	Equivalent Rewritings	132
5.3.1	Algorithm for CQ Queries, Views, and Rewritings	133
5.3.2	Finding Equivalent Rewritings for Weakly Acyclic LAV Dependencies	134
5.4	MCRs	136
5.4.1	Functional Dependencies Need Recursion	137
5.4.2	Inverse-rule Algorithm for Finding MCRs in Presence of Egds	139
5.5	Exercises	140
6	Answering Queries in Data Exchange	143
6.1	Complete Data Exchange	144
6.2	Data Exchange with Arithmetic Comparisons	146
6.2.1	Dependencies with Arithmetic Comparisons	147

6.2.2	The AC-chase	149
6.2.3	Solutions and Universal Solutions	152
6.2.4	Solutions and Conjunctive Queries	154
6.2.5	Query Answering	155
6.2.6	When the Homomorphism Property Holds—Preliminaries	158
6.2.7	Succinct AC-chase	160
6.2.8	Computing Certain Answers Using Succinct AC-chase	163
6.3	Incomplete Data Exchange	166
6.3.1	Incomplete Instances	166
6.3.2	The IDE Setting and the Corresponding CDE Settings	167
6.3.3	Computing Certain Answers in IDE Settings	168
6.4	Exercises	170
7	Answering Queries Using Views	173
7.1	Certain Answers for Queries in Presence of View Instances	173
7.1.1	Closed vs. Open World Assumption	173
7.1.2	Certain Answers vs. MCRs	174
7.2	Determinacy	178
7.2.1	Definitions and Preliminaries	178
7.2.2	Path Queries—Single View	180
7.2.3	Path Queries—CQ is Almost Complete for Rewriting	181
7.2.4	Chain Queries	184
8	XPath Queries and Views	187
8.1	XML Databases and XPath Queries	187
8.2	XPath Queries vs. Conjunctive Queries	189
8.3	XPath Query Containment and Equivalence	190
8.4	Definition of Extended Embedding between Patterns	191
8.5	Containment Test for $XP^{\{//, []\}}$ and $XP^{\{[], *\}}$	191
8.6	Extended Embedding is Not Enough to Prove Containment	192
8.7	Canonical Models	193
8.8	Containment for General Case	193
8.9	Containment and Equivalence of Union of XPath Queries	194
8.9.1	Containment Test for $XP^{\{//, *\}}$ for Single-Pattern Query	195
8.9.2	Descendant Unrollings	196
8.9.3	Containment Test for $XP^{\{//, *\}}$ for Union-of-Patterns Query	197
8.10	Rewritings	198

8.10.1	Pattern Composition	199
8.10.2	Rewritings for Query and View in $XP^{\{[],*\}}$	200
8.10.3	Rewritings that are Unions of XPath Queries	200
8.11	Conclusion and Bibliographical Notes	200
8.12	Exercises	202
9	Tree-Structured Records Queried with SQL Dialect	203
9.1	Trees as Data and as Data Types	204
9.1.1	Representing Data Types and Schemas as Trees	205
9.1.2	Instances of a Schema (Representing Data as Trees)	206
9.2	Querying Tree-Structured Data	207
9.2.1	Filter Queries	208
9.2.2	Dominance Relation	209
9.2.3	Tree-Pruning Algorithm for Filter Queries	210
9.3	Flattening	211
9.3.1	Using Flattened Data to Answer Filter Queries	214
9.4	Discussion: Tree-Pruning vs. Flattening	215
9.4.1	The Role of NULLs for Computing Legitimate Queries	217
9.5	Flattening for Linear Schemas	218
9.6	Aggregate Queries	219
9.7	Conclusion and Bibliographical Notes	221
9.7.1	Discussion on Related Work	221
10	Bibliographical Notes for Chapters 1–7	223
10.1	Query Containment	226
10.1.1	Query Containment—Set Semantics	226
10.1.2	Query Containment—Bag Semantics, Aggregation	227
10.1.3	Acyclicity	228
10.2	Query Rewriting	228
10.2.1	Binding Patterns	230
10.3	Dependencies—The Chase	230
10.4	Data Exchange	231
10.5	Other Related Work	232
10.5.1	Determinacy	232
10.5.2	More Recent Related Work	233

11	Conclusion for Chapters 1–7	235
	Bibliography	237
	Authors' Biographies	253

Preface to the First Edition

Views are used in various scenarios; some of the view-based settings have been considered in depth by the research community. The settings that we cover in this book include the following.

1. The problem of rewriting queries using views, where a set of views and a set of queries are given, and we need to find equivalent rewritings (if they exist) of the queries using these views.
2. Sometimes we cannot find equivalent rewritings but can still compute a significant part of the answer to the query. This gives rise to the problems of computing certain answers and of finding maximally contained query rewritings.
3. The picture of finding rewritings changes when we assume that the data satisfy certain constraints (dependencies). We re-examine the problem of finding rewritings for the setting in which the constraints are tuple-generating dependencies and equality-generating dependencies.
4. A closely related topic based on the same theoretical foundations is the data-exchange setting. We define the concept of certain answers and present algorithms to find them.
5. Some theoretical aspects of the more general problem of answering queries using views have also been investigated in more general and abstract settings, such as determining the (query) language fragments for which there are rewritings in the cases where the queries are determined by the views.

In order to solve problems in the above settings, we need technical tools. Thus, we provide detailed treatments of some tools, including conjunctive-query containment (with and without arithmetic comparisons and negation), the chase algorithm for reasoning about dependencies, and going beyond nonrecursive languages to find rewritings, with a discussion of Datalog.

What the book is not about: This book is not about indexes or data structures that implement the techniques considered in the exposition. Instead, this book focuses on the formal-logic perspective on the topic.

The book is written in a linear way, in the sense that each chapter depends on all the previous chapters. We have made every effort for the book to be self-contained, thus there are no substantial prerequisites. A reader familiar with the basics of the theory of database systems and knowledge of logic will move faster through the chapters. Exercises are included. More exercises,

including online exercises, and any supplementary material will be found on the [website](#)¹ of the book.

Foto Afrati and Rada Chirkova
October 2017

¹bit.ly/2zMYBHf

Preface to the Second Edition

We have added two chapters that refer to tree-like data and respective query languages. Chapter 8 presents the data model for XML documents and the XPath query language and gives query containment tests and query rewriting techniques. However, besides the relational data model and the tree-structured XML data model, many recently developed systems, to manage big data, use data models that combine relational and tree-like features. In terms of defining and using views, these models have not been investigated. However, they are mature enough to be presented in a rigorous way. Thus, Chapter 9 provides a theoretical presentation of such data model and query language as extensions of the relational model and SQL query language. In these data models, the tuples of a relation share a tree-structured schema for that relation and the query language is a dialect of SQL. The query language uses SQL-style syntax but the evaluation techniques have to be modified to fit the richer schema. The conclusions and bibliographical notes for these two chapters are included in the chapters themselves. Chapters 8 and 9 can also be read independently of the rest of the chapters in the book and independently of each other.

In the chapters of the first edition, we have corrected errors.

Foto Afrati and Rada Chirkova
January 2019

Acknowledgments

Rachel Pottinger kindly reviewed this book and provided constructive comments. Jeff Ullman contributed immensely in an elegant and concise way in the improvement of this book. We acknowledge here his many comments that have greatly improved and enhanced the text.

Foto Afrati and Rada Chirkova
October 2017