

Data Exploration Using Example-Based Methods

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.

Data Exploration Using Example-Based Methods

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

Yunyao 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, Pınar 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.

Data Exploration Using Example-Based Methods

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

ISBN: 978-3-031-00738-5 paperback

ISBN: 978-3-031-01866-4 ebook

ISBN: 978-3-031-00093-5 hardcover

DOI 10.1007/978-3-031-01866-4

A Publication in the Springer series

SYNTHESIS LECTURES ON DATA MANAGEMENT

Lecture #53

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

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

Series ISSN

Print 2153-5418 Electronic 2153-5426

Data Exploration Using Example-Based Methods

Matteo Lissandrini
Aalborg University

Davide Mottin
Aarhus University

Themis Palpanas
Paris Descartes University

Yannis Velegrakis
University of Trento

SYNTHESIS LECTURES ON DATA MANAGEMENT #53

ABSTRACT

Data usually comes in a plethora of formats and dimensions, rendering the exploration and information extraction processes challenging. Thus, being able to perform exploratory analyses in the data with the intent of having an immediate glimpse on some of the data properties is becoming crucial. Exploratory analyses should be simple enough to avoid complicate declarative languages (such as SQL) and mechanisms, and at the same time retain the flexibility and expressiveness of such languages. Recently, we have witnessed a rediscovery of the so-called *example-based methods*, in which the user, or the analyst, circumvents query languages by using examples as input. An example is a representative of the intended results, or in other words, an item from the result set. Example-based methods exploit inherent characteristics of the data to infer the results that the user has in mind, but may not be able to (easily) express. They can be useful in cases where a user is looking for information in an unfamiliar dataset, when the task is particularly challenging like finding duplicate items, or simply when they are exploring the data. In this book, we present an excursus over the main methods for exploratory analysis, with a particular focus on example-based methods. We show how that different data types require different techniques, and present algorithms that are specifically designed for relational, textual, and graph data. The book presents also the challenges and the new frontiers of machine learning in online settings which recently attracted the attention of the database community. The lecture concludes with a vision for further research and applications in this area.

KEYWORDS

search by example, data exploration, information retrieval, data management

Contents

| | | |
|----------|--|--------------|
| | Preface | xv |
| | Acknowledgments | xvii |
| 1 | Introduction | 1 |
| 1.1 | Example-Driven Exploration | 2 |
| 1.1.1 | Problem Formulation | 5 |
| 1.1.2 | Applications of Example-Based Methods | 5 |
| 1.2 | Road Map | 6 |
| | PART I Example-Based Approaches | 9 |
| 2 | Relational Data | 11 |
| 2.1 | Preliminaries | 12 |
| 2.2 | Reverse Engineering Queries (REQ) | 14 |
| 2.2.1 | Exact Reverse Engineering | 14 |
| 2.2.2 | Approximate Reverse Engineering | 20 |
| 2.3 | Schema Mapping | 25 |
| 2.3.1 | From Schema Mapping to Examples | 25 |
| 2.3.2 | Example-Driven Schema Mapping | 26 |
| 2.4 | Data Cleaning | 28 |
| 2.4.1 | Entity Matching | 29 |
| 2.4.2 | Interactive Data Repairing | 30 |
| 2.5 | Example-Based Data Exploration Systems | 33 |
| 2.6 | Summary | 34 |
| 3 | Graph Data | 37 |
| 3.1 | The Graph Data Model | 38 |
| 3.2 | Search by Example Nodes | 40 |
| 3.2.1 | Connectivity and Closeness | 40 |

| | | |
|----------|---|-----------|
| 3.2.2 | Clusters and Node Attributes | 47 |
| 3.2.3 | Similar Entity Search in Information Graphs | 49 |
| 3.3 | Reverse Engineering Queries on Graphs | 52 |
| 3.3.1 | Learning Path Queries on Graphs | 52 |
| 3.3.2 | Reverse Engineering SPARQL Queries | 55 |
| 3.4 | Search by Example Structures | 58 |
| 3.4.1 | Graph Query via Entity-Tuples | 58 |
| 3.4.2 | Queries with Example Subgraphs | 60 |
| 3.5 | Summary | 65 |
| 4 | Textual Data | 67 |
| 4.1 | Documents as Examples | 68 |
| 4.1.1 | Learning Relevance from Plain-Text | 69 |
| 4.1.2 | Modeling Networks of Document | 74 |
| 4.2 | Semi-Structured Data as Example | 78 |
| 4.2.1 | Relation Extraction | 79 |
| 4.2.2 | Incomplete Web Tables | 82 |
| 4.3 | Summary | 86 |
| 5 | Unifying Example-Based Approaches | 89 |
| 5.1 | Data Model Conversion | 89 |
| 5.2 | Seeking Relations | 92 |
| 5.2.1 | Implicit Relation | 92 |
| 5.2.2 | Explicit Relation | 93 |
| 5.3 | Entity Extraction and Matching | 94 |
| | PART II Open Research Directions | 97 |
| 6 | Online Learning | 99 |
| 6.1 | Passive Learning | 100 |
| 6.1.1 | First- and Second-Order Learning | 101 |
| 6.1.2 | Regularization | 101 |
| 6.1.3 | MindReader | 102 |
| 6.1.4 | Multi-View Learning | 103 |
| 6.2 | Active Learning | 103 |

| | | |
|----------|---|------------|
| 6.2.1 | Multi-Armed Bandits | 104 |
| 6.2.2 | Gaussian Processes | 106 |
| 6.3 | Explore-by-Example | 110 |
| 7 | The Road Ahead | 113 |
| 7.1 | Supporting Interactive Explorations | 115 |
| 7.1.1 | Query Processing | 115 |
| 7.1.2 | Automatic Result Analysis | 116 |
| 7.2 | Presenting Answers and Exploration Alternatives | 117 |
| 7.2.1 | Results Presentation | 117 |
| 7.2.2 | Generation of Exploration Alternatives | 118 |
| 7.3 | New Challenges | 119 |
| 7.3.1 | Explore Heterogeneous Data | 120 |
| 7.3.2 | Personalized Explorations | 120 |
| 7.3.3 | Exploration for Everybody | 121 |
| 8 | Conclusions | 123 |
| | Bibliography | 125 |
| | Authors' Biographies | 145 |

Preface

Exploration is one of the primordial ways to accrue knowledge about the world and its nature. It describes the act of becoming familiar with something by testing or experimenting, and at the same time it evokes the image of a traveler traversing a new territory. As we accumulate, mostly automatically, data at unprecedented volumes and speed, our datasets have become less and less familiar to us. In this context we speak of **exploratory search** as of the process of gradual discovery and understanding of the portion of the data that is pertinent to an often-times vague user's information need. Contrary to traditional search, where the desired result is well defined and the focus is on precision and performance, exploratory search usually starts from a *tentative query* that hopefully leads to answers at least partially relevant and that can provide cues about the next query. By understanding the distinction between a traditional query and an exploratory query, we can change the semantics of the user input: instead of a strict prescription of the contents of the result-set, we provide a hint of what is relevant. This shift in semantics has led to a number of methods having in common the very specific paradigm of *search by-example*. Search by-example receives as query a set of example members of the answer set. The search system then infers the entire answer set based on the given examples and any additional information provided by the underlying database.

With this book we have surveyed more than 200 research sources to highlight the main example-based techniques for relational, graph, and textual data. The book provides insights on how these example-based search systems can be employed by expert and non-expert users in retrieving the portion of the data that is relevant to their interest, while avoiding the use of complex query languages. We hope this book answers the questions and builds the necessary knowledge to those interested in constructing new data exploration systems.

Graduate students would hopefully deepen their interest in the subject and being involved in the new challenges and opportunities allowed by the powerful exploration method of search-by-example. Researchers and practitioners working in the area will probably find new insights for further improving their approaches and systems.

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

Acknowledgments

We would like to sincerely thank the authors of the referenced papers for providing us with extra support material that greatly helped the writing of this book, as well as the reviewers for their detailed and constructive comments.

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