Data Exploration Using Example-Based Methods

Synthesis Lectures on Data Management

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
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

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

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 scries
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 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

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