

Natural Language Data Management and Interfaces

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.

Natural Language Data Management and Interfaces

Yunyao Li and Davood Rafiei
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 2018

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.

Natural Language Data Management and Interfaces

Yun Yao Li and Davood Rafiei

ISBN: 978-3-031-00734-7 paperback

ISBN: 978-3-031-01862-6 ebook

ISBN: 978-3-031-00089-8 hardcover

DOI 10.1007/978-3-031-01862-6

A Publication in the Springer series

SYNTHESIS LECTURES ON DATA MANAGEMENT

Lecture #49

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

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

Series ISSN

Print 2153-5418 Electronic 2153-5426

Natural Language Data Management and Interfaces

Yunyao Li
IBM Research – Almaden

Davood Rafiei
University of Alberta

SYNTHESIS LECTURES ON DATA MANAGEMENT #49

ABSTRACT

The volume of natural language text data has been rapidly increasing over the past two decades, due to factors such as the growth of the Web, the low cost associated with publishing, and the progress on the digitization of printed texts. This growth combined with the proliferation of natural language systems for search and retrieving information provides tremendous opportunities for studying some of the areas where database systems and natural language processing systems overlap.

This book explores two interrelated and important areas of overlap: (1) managing natural language data and (2) developing natural language interfaces to databases. It presents relevant concepts and research questions, state-of-the-art methods, related systems, and research opportunities and challenges covering both areas. Relevant topics discussed on natural language data management include data models, data sources, queries, storage and indexing, and transforming natural language text. Under natural language interfaces, it presents the anatomy of these interfaces to databases, the challenges related to query understanding and query translation, and relevant aspects of user interactions. Each of the challenges is covered in a systematic way: first starting with a quick overview of the topics, followed by a comprehensive view of recent techniques that have been proposed to address the challenge along with illustrative examples. It also reviews some notable systems in details in terms of how they address different challenges and their contributions. Finally, it discusses open challenges and opportunities for natural language management and interfaces.

The goal of this book is to provide an introduction to the methods, problems, and solutions that are used in managing natural language data and building natural language interfaces to databases. It serves as a starting point for readers who are interested in pursuing additional work on these exciting topics in both academic and industrial environments.

KEYWORDS

natural language data, natural language interfaces, natural language queries, querying natural language text, semantic parsing, human computer interaction, conversational natural language interfaces

*To my husband, Huahai,
my son, Boyan,
and my parents with love*

Yunyao Li

*To my wife, Malan,
my children, Amin, Yasmin, and Ali,
and my parents with love*

Davood Rafiei

Contents

	Preface	xvii
	Acknowledgments	xix
1	Introduction	1
2	Background	5
2.1	Part-of-Speech Tagging	5
2.2	Morphological Analysis	6
2.3	Syntactic and Dependency Parsing	7
2.4	Semantic Parsing	8
2.5	Question Answering	10
2.6	Dialog System	11
3	Natural Language Data Management	15
3.1	Overview	15
3.2	Data Sources	17
3.3	Data Models	18
3.3.1	Interpreting Context as Schema	18
3.3.2	Mapping Content to More Formal Models	18
3.4	Queries	20
3.4.1	Boolean Keyword Queries	20
3.4.2	Grammar-Based Searches	21
3.4.3	Text Pattern Queries	21
3.4.4	Tree Pattern Queries	23
3.4.5	Combining Text and Tree Pattern Queries	24
3.4.6	Summary	25
3.5	Indexing Natural Language Text	25
3.5.1	Indexing for Text Pattern Queries	25
3.5.2	Indexing for Tree Pattern Queries	28
3.5.3	Summary	32
3.6	Transforming Natural Language Text	32

3.6.1	Meaning Representation	33
3.6.2	Meaning of Words	35
3.6.3	Computing Word Semantics	38
3.6.4	Meaning of Sentences	40
3.6.5	Information Extraction	42
3.6.6	Entity Linking	43
3.6.7	Summary and Discussions	45
3.7	Summary	46
4	Natural Language Interfaces to Databases	47
4.1	Overview	47
4.1.1	Anatomy	48
4.1.2	Challenges	48
4.1.3	Summary	49
4.2	Query Understanding	49
4.2.1	Scope	50
4.2.2	Stateless vs. Stateful	51
4.2.3	Parser Error Handling	54
4.3	Query Translation	55
4.3.1	Bridging the Semantic Gap	56
4.3.2	Query Construction	63
4.4	User Interactions	66
4.4.1	Design Considerations	66
4.4.2	User Interaction Models	66
4.4.3	Stateless vs. Stateful	67
4.5	Notable Systems	68
4.5.1	PRECISE	68
4.5.2	NLPQC	71
4.5.3	NaLIX	73
4.5.4	FREyA	79
4.5.5	NaLIR	82
4.5.6	NL ₂ CM	84
4.5.7	ATHANA	87
4.5.8	SQLizer	94
4.5.9	Seq2SQL	98
4.5.10	Summary	101
4.6	Relationship to Other Areas	102

4.6.1	Relationship to Question Answering	102
4.6.2	Relationship to Semantic Parsing	102
4.7	Summary	105
5	Open Challenges and Opportunities	107
5.1	Resolving References	107
5.2	Understanding Natural Language Text and Queries	108
5.3	Mobile Natural Language Data Management	108
5.4	Multilingual/Cross-Lingual Support	108
5.5	Evaluation	109
6	Conclusions	111
	Bibliography	113
	Authors' Biographies	131
	Index	133

Preface

Natural languages are languages developed naturally by human through use and repetition. They are central to almost all human activities. In today's digital world, a large portion of data that is stored and exchanged is in natural languages. These languages also play a growing role in our daily interactions with machines with the popularization of voice-based interfaces such as self-driven cars and virtual personal assistants. Allowing "casual users" to employ their native languages today has implications for both communicating with databases and storing and retrieving data in the form of natural languages.

This book grew out of our passion for the two interrelated topics in the intersection of database systems and natural language processing: managing natural languages and building natural language interfaces to databases. Despite the commonality of the issues in understanding natural languages and dealing with ambiguities, each area offers some challenges of its own. In the former, the structure of the data is described informally in a natural language but the queries are more formal. In the latter, data is described more formally (e.g., in a relational database) but the queries are informally expressed in a natural language.

This goal of this book is to provide a unified view of both topics, with overlapping areas discussed once and/or cross-referenced. This book takes a structured approach to present a comprehensive survey of all important research problems and their key sub-problems and the latest development in the related fields. It also bridges the gap between everything-is-a-relation and everything-is-a-text cultures, highlighting where each culture shines and how it contributes to an integrated solution.

This book is suitable for database students, researchers, and developers who are interested in different aspects of managing natural language data and developing natural language interfaces to databases. It will also provide students, researchers, and practitioners in other related areas (such as natural language processing, question answering, information retrieval, data mining, and machine learning) with database principles and techniques that may be applicable to related problems in those areas.

The book may be used within various courses at graduate and undergraduate levels, as a starting point to the literature. A course covering natural language interfaces to databases may discuss Sections 4.1–4.4 for the main components, their functions and challenges, and one or more of the systems in Section 4.5, as relevant, for more details. A course covering querying and indexing natural language text may discuss Sections 3.3–3.5 and maybe Section 3.6. In both cases, any other section may be covered as relevant or applicable. Section 3.6 may also be covered within a course on linked data and semantic search. The background section may be skipped for those familiar with common natural language processing techniques. The book grew

xviii PREFACE

out of a three-hour tutorial on the same subject [Li and Rafiei, 2017], given by the authors at SIGMOD 2017. The slides used in the tutorial are available online and can be easily incorporated into courses.¹

Yunyao Li and Davood Rafiei
July 2018

¹https://webdocs.cs.ualberta.ca/~drafie/papers/SIGMOD2017tutorial_LR.pdf and <https://www.slideshare.net/YunyaoLi/natural-language-data-management-and-interfaces-recent-development-and-open-challenges>

Acknowledgments

This book is made possible with the help and support of a number of people. H.V. Jagadish invited us to write this book after our SIGMOD 2017 tutorial, and Diane Cerra pushed us to stay within our timeline. We wish to thank both for their help and support. Joint work and discussions with our students and colleagues over the years helped shape the book. In particular, Davood wishes to thank Pirooz Chubak, Haobin Li, Dekang Lin, and Ehsan Kamalloo, and Yunyao wishes to thank Alan Akbik, Chris Baik, Ishan Chaudhuri, Laura Chiticariu, Ronald Fagin, Laura Haas, H.V. Jagadish, Benny Kimelfeld, Rajasekar Krishnamurthy, Sriram Raghavan, Lucian Popa, Frederick Reiss, Satinder P. Singh, Shivakumar Vaithyanathan, Huahai Yang, and Huaiyu Zhu. We are also grateful to our reviewers (Daniel Deutch, Sourav Bhowmick, and Laszlo Kovacs) who, despite their busy schedules, carefully read the initial draft of this book and provided detailed and constructive comments. We have taken each and every one of them into consideration while improving the book. Finally, we also wish to acknowledge our funding sources. Davood Rafiei's research is supported by the Natural Sciences and Engineering Research Council of Canada, and Yunyao Li's research is supported by IBM Research.

Yunyao Li and Davood Rafiei
July 2018