Natural Language Data Management and Interfaces

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.

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

Natural Language Data Management and Interfaces

Yunyao 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

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

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

[1] https://webdocs.cs.ualberta.ca/~drafiei/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