Managing Data From Knowledge Bases: Querying and Extraction

Wei Emma Zhang • Quan Z. Sheng

Managing Data From Knowledge Bases: Querying and Extraction

Springer

Wei Emma Zhang
Department of Computing
Macquarie University
Sydney, NSW, Australia

Quan Z. Sheng
Department of Computing
Macquarie University
Sydney, NSW, Australia

ISBN 978-3-319-94934-5 ISBN 978-3-319-94935-2 (eBook)
https://doi.org/10.1007/978-3-319-94935-2

Library of Congress Control Number: 2018950437

Printed on acid-free paper

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Foreword

Knowledge bases (KBs) are the most essential components in realizing semantic computing for better human-machine interaction experiences. Knowledge bases supply facts and relationships for use in computation by machines. This can facilitate artificial intelligence (AI) tools with the ability to reason and explain. Over the years, knowledge base has been receiving much attention, both from academia and industry, as a resource for providing knowledge, an auxiliary tool for facilitating the searching on search engines, and an expert system for helping in decision making.

Knowledge available for improving computations by AI tools has grown to become quite large, which presents a number of technical challenges including efficient knowledge retrieval and automatic knowledge base construction. Among the books on the market that cover various challenges related to KBs, this book presents one of the rare attempts to present innovative solutions for the knowledge extraction and querying in knowledge bases.

These topics are under the umbrella of extracting knowledge from unstructured data for the effective construction of knowledge bases and querying knowledge bases based on a learning-based cache framework. The book overviews key findings from the authors' intensive research experience in analyzing data from different knowledge sources for knowledge base queries and knowledge base construction. The extensive references included in this book will help the interested readers find out more information on the discussed topics.

I am happy to commend the authors for their outstanding accomplishment and to inform the readers that they are looking at an authoritative piece of work in the vibrant and rapidly expanding field of knowledge extraction and discovery. This book is a valuable resource for everyone interested in the topics this book covers in depth.

Dayton, OH, USA
April 2018

Amit Sheth

Preface

Semantic Web is a paradigm that publishes and retrieves knowledge on the Web in a semantically structured way. Knowledge base (KB) is one of the most essential components in realizing the idea of Semantic Web as it provides facts and relationships that can be automatically understood, interpreted, and deduced by machines (e.g., programmatic software). Recently, knowledge base has gained momentum in providing accurate, expert, and multidisciplinary knowledge to the society. While it is well understood that knowledge base offers numerous opportunities and benefits, it also presents significant technical challenges. Among them, effective and efficient knowledge extraction and retrieval are two fundamental challenges facing the research community and industry today.

In this book, we first address the research issues and explore the principles and techniques of the challenging topics. Then we solve the raised research issues by developing a series of methodologies. More specifically, we study the query optimization and tackle the query performance prediction for knowledge retrieval. We also handle unstructured data processing and data clustering for knowledge extraction. To optimize the queries issued through interfaces against knowledge bases, we propose a cache-based optimization layer between consumers and the querying interface to facilitate the querying and solve the latency issue. The cache depends on a novel learning method that considers the querying patterns from individual's historical queries without having knowledge of the backing systems of the knowledge base. To predict the query performance for appropriate query scheduling, we examine the queries' structural and syntactical features and apply multiple widely adopted prediction models. Our feature modeling approach eschews the knowledge requirement on both the querying languages and system. To extract knowledge from unstructured Web sources, we examine two kinds of Web sources containing unstructured data: the source code from Web repositories and the posts in programming question-answering communities. We use natural language processing techniques to pre-process the source codes and obtain the natural language elements. Then we apply traditional knowledge extraction techniques to extract knowledge. For the data from programming question-answering communities, we make the attempt towards building programming knowledge base by starting with paraphrase identification problem and develop novel features to accurately identify duplicate posts. For domain-specific knowledge extraction, we propose to use clustering technique to separate knowledge into different groups. We focus on developing a new clustering algorithm that uses manifold constraint in the optimization task and achieves fast and accurate performance. For each of model and approach presented in this book, we have conducted extensive experiments to evaluate it using either public dataset or synthetic data we generated. We also discuss some open research directions at the end of this book.

Sydney, NSW, Australia
April 2018

Wei Emma Zhang
Quan Z. Sheng

Acknowledgments

I would like to express my gratitude to Prof. Michael Sheng, Prof. Mingkui Tan, and A/Prof. Kerry Taylor for their help and guidance. I owe a huge debt to my parents, my husband, and my prince for their support and sacrifice. They have been always there for me whenever I needed them. Without them, I could not be successful at any point of time.

Wei Emma Zhang

To my mum for her love.

Quan Z. Sheng

Contents