

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

© Springer International Publishing AG, part of Springer Nature 2018

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

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

1	Introduction	1
1.1	Overview of Knowledge Bases	2
1.2	Overview of Knowledge Extraction in Knowledge Bases	4
1.2.1	Extraction Techniques Overview	4
1.2.2	Representation Models Overview	7
1.3	Overview of Knowledge Bases Question Answering	10
1.3.1	Question Answering in Curated KBs	10
1.3.2	Question Answering in Open KBs	14
1.4	Research Issues in Querying and Extracting Knowledge Bases	14
1.4.1	An Architecture for Knowledge Base Management	15
1.4.2	Our Contributions	16
1.5	Book Organization	18
2	Cache Based Optimization for Querying Curated Knowledge Bases	19
2.1	Design Overview	19
2.2	The SPARQL Endpoint Cache Framework	23
2.2.1	Query Distance Calculation	24
2.2.2	Feature Modelling	28
2.2.3	Suggesting and Prefetching Similar Queries	29
2.2.4	Caching and Replacement	31
2.3	Experimental Evaluations and Discussions	34
2.3.1	Setup	34
2.3.2	Analysis of Real-World SPARQL Queries	35
2.3.3	Performance of Cache Replacement Algorithm	37
2.3.4	Comparison of Feature Modelling Approaches	39
2.3.5	Performance Comparison with the State-of-the-Art	42
2.3.6	Discussions	43
2.4	Related Work	44
2.4.1	Semantic Caching	44
2.4.2	Query Suggestion	45
2.5	Summary	46

3	Query Performance Prediction on Knowledge Base	47
3.1	Design Overview	47
3.1.1	Motivation	48
3.1.2	Challenges	49
3.1.3	Prediction Approach Overview	50
3.2	Preliminaries	52
3.2.1	Multiple Regression	52
3.2.2	Dimension Reduction	53
3.3	Feature Modelling for Queries	53
3.3.1	Algebra Features	53
3.3.2	BGP Features	55
3.3.3	Hybrid Features	55
3.4	Predicting Query Performance	56
3.4.1	Predictive Models	56
3.4.2	Two-Step Prediction	57
3.5	Experimental Evaluation and Discussion	58
3.5.1	Setup	58
3.5.2	Prediction Models Comparison	59
3.5.3	Feature Modelling Comparison	60
3.5.4	Comparison of Different Weighting Schemes in k -NN Regression	62
3.5.5	Performance of Two-Step Prediction	63
3.5.6	Comparison to State-of-the-Art	64
3.6	Discussions	65
3.7	Related Work	66
3.7.1	Query Performance Prediction via Machine Learning Algorithms	66
3.7.2	SPARQL Query Optimization	67
3.8	Summary	67
4	An Efficient Knowledge Clustering Algorithm	69
4.1	Overview of Clustering with Non-negative Matrix Factorization	69
4.2	Orthogonal Non-negative Matrix Factorization Over Stiefel Manifold	71
4.2.1	Notations	71
4.2.2	Optimization on Stiefel Manifold	71
4.2.3	Update U via NRCG	73
4.2.4	Update V	76
4.2.5	Convergence Analysis	78
4.3	Experimental Evaluation	78
4.3.1	Implementation Details	79
4.3.2	Data Sets	79
4.3.3	Metrics	81
4.3.4	Results	82
4.4	Related Works	87
4.5	Summary	88

- 5 Knowledge Extraction from Unstructured Data on the Web** 89
 - 5.1 Design Overview 89
 - 5.2 Source Code Topics Extraction via Topic Model and Words Embedding 91
 - 5.2.1 Data Pre-processing 92
 - 5.2.2 Topic Extraction 92
 - 5.2.3 The Coherence Measurement 95
 - 5.2.4 Automated Terms Selection for Topic Extraction 96
 - 5.3 Experimental Evaluation 97
 - 5.3.1 Setup 97
 - 5.3.2 Results 98
 - 5.4 Related Works 101
 - 5.5 Summary 102
- 6 Building Knowledge Bases from Unstructured Data on the Web** 103
 - 6.1 Design Overview 103
 - 6.2 Prototype of Knowledge Extraction from Programming Question Answering Communities 105
 - 6.2.1 Question Extraction 106
 - 6.2.2 Answer and Tags Extraction 106
 - 6.2.3 Triple Generation 106
 - 6.3 Detecting Duplicate Posts in Programming QA Communities 107
 - 6.3.1 Pre-processing 107
 - 6.3.2 Feature Modelling 108
 - 6.3.3 Binary Classification 111
 - 6.4 Experimental Evaluation and Discussions 112
 - 6.4.1 Setup 112
 - 6.4.2 Results 114
 - 6.4.3 Discussions 119
 - 6.5 Related Work 120
 - 6.5.1 Question Retrieval from QA Communities 120
 - 6.5.2 Mining PCQA Websites 121
 - 6.6 Summary 121
- 7 Conclusion** 123
 - 7.1 Summary 123
 - 7.2 Future Directions 125
- References** 127