

Question Answering for the Curated Web

Tasks and Methods in QA over Knowledge Bases
and Text Collections

Synthesis Lectures on Information Concepts, Retrieval, and Services

Editor

Gary Marchionini, *University of North Carolina, Chapel Hill*

Synthesis Lectures on Information Concepts, Retrieval, and Services publishes short books on topics pertaining to information science and applications of technology to information discovery, production, distribution, and management. Potential topics include: data models, indexing theory and algorithms, classification, information architecture, information economics, privacy and identity, scholarly communication, bibliometrics and webometrics, personal information management, human information behavior, digital libraries, archives and preservation, cultural informatics, information retrieval evaluation, data fusion, relevance feedback, recommendation systems, question answering, natural language processing for retrieval, text summarization, multimedia retrieval, multilingual retrieval, and exploratory search.

Question Answering for the Curated Web: Tasks and Methods in QA over Knowledge Bases and Text Collections

Rishiraj Saha Roy and Avishek Anand
2021

Third Space, Information Sharing, and Participatory Design

Preben Hansen, Ina Fourie, and Anika Meyer
2021

Trustworthy Communications and Complete Genealogies: Unifying Ancestries for a Genealogical History of the Modern World

Chirag Shah and Ryen W. White
2021

Trustworthy Communications and Complete Genealogies: Unifying Ancestries for a Genealogical History of the Modern World

Reagan W. Moore
2021

Word Association Thematic Analysis: A Social Media Text Exploration Strategy

Michael Thelwall

2021

Simulating Information Retrieval Test Collections

David Hawking, Bodo Billerbeck, Paul Thomas, and Nick Craswell

2020

Automatic Disambiguation of Author Names in Bibliographic Repositories

Anderson A. Ferreira, Marcos André Gonçalves, and Alberto H. F. Laender

2020

Compatibility Modeling: Data and Knowledge Applications for Clothing Matching

Xuemeng Song, Liqiang Nie, and Yinglong Wang

2019

Video Structure Meaning

Brian O'Connor and Rich Anderson

2019

Interactive IR User Study Design, Evaluation, and Reporting

Jiqun Liu and Chirag Shah

2019

The Practice of Crowdsourcing

Omar Alonso

2019

Predicting Information Retrieval Performance

Robert M. Losee

2018

Framing Privacy in Digital Collections with Ethical Decision Making

Virginia Dressler

2018

Mobile Search Behaviors: An In-depth Analysis Based on Contexts, APPs, and Devices

Dan Wu and Shaobo Liang

2018

Images in Social Media: Categorization and Organization of Images and Their Collections

Susanne Ørnager and Haakon Lund

2018

Exploring Context in Information Behavior: Seeker, Situation, Surroundings, and Shared Identities

Naresh Kumar Agarwal
2017

Researching Serendipity in Digital Information Environments

Lori McCay-Peet and Elaine G. Toms
2017

Social Monitoring for Public Health

Michael J. Paul and Mark Dredze
2017

Digital Libraries for Cultural Heritage: Development, Outcomes, and Challenges from European Perspectives

Tatjana Aparac-Jelšić
2017

iRODS Primer 2: Integrated Rule-Oriented Data System

Hao Xu, Terrell Russell, Jason Coposky, Arcot Rajasekar, Reagan Moore, Antoine de Torcy, Michael Wan, Wayne Shroeder, and Sheau-Yen Chen
2017

Information Architecture: The Design and Integration of Information Spaces, Second Edition

Wei Ding, Xia Lin, and Michael Zarro
2017

Fuzzy Information Retrieval

Donald H. Kraft and Erin Colvin
2017

Quantifying Research Integrity

Michael Seadle
2016

Incidental Exposure to Online News

Borchuluun Yadamsuren and Sanda Erdelez
2016

Web Indicators for Research Evaluation: A Practical Guide

Michael Thelwall
2016

Trustworthy Policies for Distributed Repositories

Reagan W. Moore, Hao Xu, Mike Conway, Arcot Rajasekar, Jon Crabtree, and Helen Tibbo
2016

The Notion of Relevance in Information Science: Everybody knows what relevance is.
But, what is it really?

Tefko Saracevic
2016

Dynamic Information Retrieval Modeling

Grace Hui Yang, Marc Sloan, and Jun Wang
2016

Learning from Multiple Social Networks

Liqiang Nie, Xuemeng Song, and Tat-Seng Chua
2016

Scholarly Collaboration on the Academic Social Web

Daqing He and Wei Jeng
2016

Scalability Challenges in Web Search Engines

B. Barla Cambazoglu and Ricardo Baeza-Yates
2015

Social Informatics Evolving

Pnina Fichman, Madelyn R. Sanfilippo, and Howard Rosenbaum
2015

On the Efficient Determination of Most Near Neighbors: Horseshoes, Hand Grenades,
Web Search and Other Situations When Close Is Close Enough, Second Edition

Mark S. Manasse
2015

Building a Better World with Our Information: The Future of Personal Information
Management, Part 3

William Jones
2015

Click Models for Web Search

Aleksandr Chuklin, Ilya Markov, and Maarten de Rijke
2015

Information Communication

Feicheng Ma
2015

Social Media and Library Services

Lorri Mon
2015

Analysis and Visualization of Citation Networks

Dangzhi Zhao and Andreas Strotmann

2015

The Taxobook: Applications, Implementation, and Integration in Search: Part 3 of a 3-Part Series

Marjorie M.K. Hlava

2014

The Taxobook: Principles and Practices of Building Taxonomies, Part 2 of a 3-Part Series

Marjorie M.K. Hlava

2014

Measuring User Engagement

Mounia Lalmas, Heather O'Brien, and Elad Yom-Tov

2014

The Taxobook: History, Theories, and Concepts of Knowledge Organization, Part 1 of a 3-Part Series

Marjorie M.K. Hlava

2014

Children's Internet Search: Using Roles to Understand Children's Search Behavior

Elizabeth Foss and Allison Druin

2014

Digital Library Technologies: Complex Objects, Annotation, Ontologies, Classification, Extraction, and Security

Edward A. Fox and Ricardo da Silva Torres

2014

Digital Libraries Applications: CBIR, Education, Social Networks, eScience/Simulation, and GIS

Edward A. Fox and Jonathan P. Leidig

2014

Information and Human Values

Kenneth R. Fleischmann

2013

Multiculturalism and Information and Communication Technology

Pnina Fichman and Madelyn R. Sanfilippo

2013

Transforming Technologies to Manage Our Information: The Future of Personal Information Management, Part II

William Jones

2013

Designing for Digital Reading

Jennifer Pearson, George Buchanan, and Harold Thimbleby

2013

Information Retrieval Models: Foundations and Relationships

Thomas Roelleke

2013

Key Issues Regarding Digital Libraries: Evaluation and Integration

Rao Shen, Marcos Andre Goncalves, and Edward A. Fox

2013

Visual Information Retrieval using Java and LIRE

Mathias Lux and Oge Marques

2013

On the Efficient Determination of Most Near Neighbors: Horseshoes, Hand Grenades, Web Search and Other Situations When Close is Close Enough

Mark S. Manasse

2012

The Answer Machine

Susan E. Feldman

2012

Theoretical Foundations for Digital Libraries: The 5S (Societies, Scenarios, Spaces, Structures, Streams) Approach

Edward A. Fox, Marcos André Gonçalves, and Rao Shen

2012

The Future of Personal Information Management, Part I: Our Information, Always and Forever

William Jones

2012

Search User Interface Design

Max L. Wilson

2011

Information Retrieval Evaluation

Donna Harman

2011

Knowledge Management (KM) Processes in Organizations: Theoretical Foundations and Practice

Claire R. McInerney and Michael E. D. Koenig
2011

Search-Based Applications: At the Confluence of Search and Database Technologies

Gregory Grefenstette and Laura Wilber
2010

Information Concepts: From Books to Cyberspace Identities

Gary Marchionini
2010

Estimating the Query Difficulty for Information Retrieval

David Carmel and Elad Yom-Tov
2010

iRODS Primer: Integrated Rule-Oriented Data System

Arcot Rajasekar, Reagan Moore, Chien-Yi Hou, Christopher A. Lee, Richard Marciano, Antoine de Torcy, Michael Wan, Wayne Schroeder, Sheau-Yen Chen, Lucas Gilbert, Paul Tooby, and Bing Zhu
2010

Collaborative Web Search: Who, What, Where, When, and Why

Meredith Ringel Morris and Jaime Teevan
2009

Multimedia Information Retrieval

Stefan Rüger
2009

Online Multiplayer Games

William Sims Bainbridge
2009

Information Architecture: The Design and Integration of Information Spaces

Wei Ding and Xia Lin
2009

Reading and Writing the Electronic Book

Catherine C. Marshall
2009

Hypermedia Genes: An Evolutionary Perspective on Concepts, Models, and Architectures

Nuno M. Guimarães and Luís M. Carrico
2009

Understanding User-Web Interactions via Web Analytics

Bernard J. (Jim) Jansen

2009

XML Retrieval

Mounia Lalmas

2009

Faceted Search

Daniel Tunkelang

2009

Introduction to Webometrics: Quantitative Web Research for the Social Sciences

Michael Thelwall

2009

Exploratory Search: Beyond the Query-Response Paradigm

Ryen W. White and Resa A. Roth

2009

New Concepts in Digital Reference

R. David Lankes

2009

Automated Metadata in Multimedia Information Systems: Creation, Refinement, Use in Surrogates, and Evaluation

Michael G. Christel

2009

© Springer Nature Switzerland AG 2022

Reprint of original edition © Morgan & Claypool 2022

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.

Question Answering for the Curated Web: Tasks and Methods in QA over Knowledge Bases
and Text Collections

Rishiraj Saha Roy and Avishek Anand

ISBN: 978-3-031-79511-4 paperback

ISBN: 978-3-031-79512-1 ebook

ISBN: 978-3-031-79513-8 hardcover

DOI 10.1007/978-3-031-79512-1

A Publication in the Springer series

SYNTHESIS LECTURES ON INFORMATION CONCEPTS, RETRIEVAL, AND SERVICES

Lecture #76

Series Editor: Gary Marchionini, *University of North Carolina, Chapel Hill*

Series ISSN

Print 1947-945X Electronic 1947-9468

Question Answering for the Curated Web

Tasks and Methods in QA over Knowledge Bases
and Text Collections

Rishiraj Saha Roy

Max Planck Institute for Informatics, Saarbruecken, Germany

Avishek Anand

Leibniz University, Hannover, Germany

*SYNTHESIS LECTURES ON INFORMATION CONCEPTS, RETRIEVAL,
AND SERVICES #76*

ABSTRACT

Question answering (QA) systems on the Web try to provide crisp answers to information needs posed in natural language, replacing the traditional ranked list of documents. QA, posing a multitude of research challenges, has emerged as one of the most actively investigated topics in information retrieval, natural language processing, and the artificial intelligence communities today. The flip side of such diverse and active interest is that publications are highly fragmented across several venues in the above communities, making it very difficult for new entrants to the field to get a good overview of the topic. Through this book, we make an attempt towards mitigating the above problem by providing an overview of the state-of-the-art in question answering. We cover the twin paradigms of curated Web sources used in QA tasks – trusted text collections like Wikipedia, and objective information distilled into large-scale knowledge bases. We discuss distinct methodologies that have been applied to solve the QA problem in both these paradigms, using instantiations of recent systems for illustration. We begin with an overview of the problem setup and evaluation, cover notable sub-topics like open-domain, multi-hop, and conversational QA in depth, and conclude with key insights and emerging topics. We believe that this resource is a valuable contribution towards a unified view on QA, helping graduate students and researchers planning to work on this topic in the near future.

KEYWORDS

question answering, knowledge bases, text collections

Dedicated to Amrita, Hia, Adi, Baba, and Didi.—Rishiraj

Dedicated to Megha, Ananya, Papa, and Mami.—Avishek

Contents

	Preface	xix
	Acknowledgments	xxi
1	Introduction	1
1.1	Motivation	1
1.2	Perspectives	2
1.3	Relevance to the Community	3
1.4	Intended Audience	3
1.5	Scope of QA in this Book	4
	 PART I QA over Knowledge Bases	 7
2	Setup	9
2.1	Introducing Knowledge Bases	9
2.1.1	Basic Concepts	9
2.1.2	Storing Contextual Information	10
2.1.3	Wikidata as an Active KB	12
2.2	QA over Knowledge Bases	13
2.2.1	Basic Concepts	13
2.2.2	Generating Answers	14
2.2.3	Evaluating KB-QA Systems	16
2.3	Organization	20
3	Getting Started with Simple Questions	23
3.1	Introducing Simple QA	23
3.2	Answering with Templates	25
3.3	Answering with KG Embeddings	32
3.4	Wrapping up Simple QA	35

4	Complex Question Answering	37
4.1	Introducing Complex QA	37
4.2	Answering with Staged Graphs	40
4.3	Answering with Compact Subgraphs	43
4.4	Answering with Belief Propagation	47
4.5	Wrapping up Complex QA	49
5	Answering over Heterogeneous Sources	53
5.1	Introducing Heterogeneous QA	53
5.2	Answering with Early Fusion	54
5.3	Answering with Unified Representations	57
5.3.1	Unifying with SPO Triples	57
5.3.2	Unifying with Verbalizations	60
5.4	Wrapping up Heterogeneous QA	62
6	Conversational Question Answering	65
6.1	Introducing Conversational QA	65
6.2	Answering with Graph Expansion	66
6.3	Answering with Seq2seq Models	69
6.4	Answering with Reformulations	73
6.5	Wrapping up Conversational QA	75
7	Part I: Summary and Insights	77
7.1	Mapping Systems to Methodologies	77
7.2	Structured Queries: With or Without?	78
7.3	Quad Chart for White Space Analysis	79
7.4	Deploying a KB-QA System Quickly	81
	 PART II QA over Text Collections	 83
8	Setup	85
8.1	Question Type	85
8.2	Answer Type	86
8.3	Context Type	88
8.4	Evaluation	89

8.5	Benchmarks	90
8.6	Organization	93
9	Reading Comprehension	95
9.1	Introducing Reading Comprehension	95
9.1.1	A Deep Learning Approach	95
9.1.2	The Reading Comprehension Pipeline	96
9.2	Token Representation	96
9.3	Input Representation	99
9.4	Interaction Modeling	99
9.5	Attention Mechanism	100
9.6	Answer Modeling	102
9.7	The DrQA Reader	103
9.7.1	The Transformer Architecture	105
9.8	BERT and Transfer Learning	107
9.8.1	The BERT Model	107
9.8.2	Pre-Training BERT	107
9.8.3	Fine-Tuning BERT for QA	108
9.9	Wrapping up Reading Comprehension	108
10	Open-Domain Question Answering	111
10.1	Introducing Open-Domain QA	111
10.2	The Retriever-Reader Framework	112
10.2.1	Challenges for Retrieve-and-Read Frameworks	113
10.3	Learning with Distant Supervision	114
10.3.1	Relaxed vs. Restrictive Distant Supervision	114
10.4	Answer Generation	115
10.5	Parameterized Retriever	116
10.5.1	Dual-Encoder Models	116
10.5.2	Retrieval with Parameterized Retrievers	117
10.5.3	Nearest Neighbor Retrieval	117
10.5.4	Training Dual-Encoder Models	118
10.5.5	Pre-Training for QA	119
10.6	Wrapping up Open-Domain QA	120

11	Multi-Hop Question Answering	121
11.1	Introducing Multi-Hop QA	121
11.2	Multi-Turn Retrieval	122
11.3	Reasoning Paths	124
11.4	Treating Text as a Knowledge Base	126
11.5	Wrapping up Multi-Hop QA	128
12	Conversational Question Answering	129
12.1	Introducing Conversational QA	129
12.1.1	Challenges in Conversational QA	129
12.2	Datasets in Conversational QA	131
12.3	The Pointer-Generator Model	132
12.3.1	Answer Extraction	132
12.3.2	Generative Answers Using PGNet	133
12.4	Flow Models	134
12.5	Wrapping up Conversational QA	136
13	Part II: Summary and Insights	137
13.1	Unsupervised Embeddings and Transfer Learning	137
13.2	Utility of a Parametric Retriever	138
13.3	Question and Passage Understanding	138
13.4	Simple Design Choices	139
13.5	Deploying a Text-QA System Quickly	140
13.6	Requirements for Commercial QA Systems	140
14	Open Directions	143
	References	147
	Authors' Biographies	171

Preface

Receiving direct answers from Web search engines and personal assistants has become an integral part of a user's digital experience today. The goal of this book is to satisfy curious minds about this hot topic and help newcomers get started into their own research on QA systems over structured and unstructured sources on the Web. The book is not meant to be a comprehensive survey of the vast body of work on question answering, but rather as a direct initiator to the state-of-the-art in the field of QA for the curated Web. As a result, it has a bias towards very recent works published at prominent venues in the IR, NLP, or AI communities. The book skips the detailed pipelines in individual research papers in favor of abstractions from different families of methods under a unified terminology. Within each sub-topic, we try to cover distinct methodologies using specific system instantiations as support. Given a question and a knowledge source, the goal is to tease apart the embellishments and formal notation and get across the core process of deriving the answer. The book closely follows the SIGIR 2020 tutorial by the authors on the same topic.

Rishiraj Saha Roy and Avishek Anand
August 2021

Acknowledgments

We thank authors of individual research papers who provided us with additional content like slide decks of talks, that helped us understand their ideas better. We are grateful to the anonymous reviewers of this book, whose critical comments helped improve the manuscript significantly. We express our gratitude to our mentor Prof. Dr. Gerhard Weikum for constructive feedback at various stages of this book. Rishiraj extends thanks to his students Magdalena Kaiser and Philipp Christmann for suggesting useful edits on various parts of the book. Avishek is also grateful to his amazing colleagues and students—Lijun Lyu, Jurek Leonhardt, Maximillian Idahl, Florian Knauf, Thorben Funke, Megha Khosla, Soumyadeep, and Koustav Rudra—who provided valuable comments and ideas on the manuscript. We thank Morgan & Claypool for providing us with this opportunity of publishing our contribution as an instalment of the Synthesis Lectures Series. Finally, most of this book was written during the long lockdown period imposed due to the corona pandemic: it goes without saying that this manuscript could simply not have been possible without the continued support of the authors' family members.

Rishiraj Saha Roy and Avishek Anand
August 2021