

The Information Retrieval Series

Krisztian Balog

Entity- Oriented Search



Springer Open

Series Editors

ChengXiang Zhai

Maarten de Rijke

Editorial Board

Nicholas J. Belkin

Charles Clarke

Diane Kelly

Fabrizio Sebastiani

More information about this series at <http://www.springer.com/series/6128>

Krisztian Balog

Entity-Oriented Search

Krisztian Balog
University of Stavanger
Stavanger, Norway



ISSN 1387-5264

The Information Retrieval Series

ISBN 978-3-319-93933-9

ISBN 978-3-319-93935-3 (eBook)

<https://doi.org/10.1007/978-3-319-93935-3>

Library of Congress Control Number: 2018946540

© The Editor(s) (if applicable) and the Author(s) 2018, This book is an open access publication.

Open Access This book is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this book are included in the book's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the book's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

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

Szüleimnek

Preface

I have not yet reached my goal. . . But I forget what is behind, and I struggle for what is ahead. I run toward the goal, so I can win the prize of being called to heaven. This is the prize God offers because of what Christ Jesus has done.

(Philippians 3:12–14, CEV)

The idea of writing this book stemmed from a series of tutorials that I gave with colleagues on “entity linking and retrieval for semantic search.” There was no single text on this topic that would cover all the material that I wished to introduce to someone who is new to this field. With this book, I set out to fill that gap. I hope that by making the book open access, many will be able to use it and benefit from it.

For me, writing this book, in many ways, was like running a marathon. No one forced me to do it, yet I thought that—for some reason—it’d be a good idea to challenge myself to do it. Then, along the way, there comes inevitably a point where one asks: Why am I doing this to myself? But then, in the end, crossing the finish line certainly feels like an accomplishment. In time, this experience might even be remembered as if it was a walk in the park.¹ In any case, it was a good run.

I wish to express my gratitude to a number of people who played a role in making this book happen. First of all, I would like to thank Ralf Gerstner, executive editor for Computer Science at Springer, for seeing me through to the successful completion of this book and for always being a gentleman when it came to my deadline extension requests. I also want to thank the Information Retrieval Series editors Maarten de Rijke and ChengXiang Zhai for the comments on my book proposal.

A very special thanks to Jamie Callan and to anonymous Reviewer #2 for reviewing the book and for making numerous valuable suggestions for improvements.

The following colleagues provided feedback on drafts of specific chapters at various stages of completion, and I would like to thank them for their insightful comments: Marek Ciglan, Arjen de Vries, Kalervo Järvelin, Miguel Martinez, Edgar

¹Note to self: No, it wasn’t.

Meij, Kjetil Nørvåg, Doug Oard, Heri Ramampiaro, Ralf Schenkel, Alberto Tonon, and Chenyan Xiong.

I want to thank Edgar Meij and Daan Odijk for the collaboration on the entity linking and retrieval tutorials, which planted the idea of this book. Working with you was always easy, enjoyable, and fun. My gratitude goes to all my co-authors for the joint work that contributed to the material that is presented in this book.

I am especially grateful to the Department of Electrical Engineering and Computer Science at the University of Stavanger for providing a pleasant work environment, where I could devote a substantial amount of time to writing this book.

I would like to thank my PhD students for giving me their honest opinion and offering constructive criticism on drafts of the book. They are, in gender-first-then-alphabetical order: Faegheh Hasibi, Jan Benetka, Heng Ding, Darío Garigliotti, Trond Linjordet, and Shuo Zhang. Special thanks, in addition, to Faegheh for the thorough checking of technical details and for suggestions on the organization of the material; to Darío for tidying up my references; to Jan for prettifying the figures and illustrations; to Trond for injecting entropy and for the careful proofreading and numerous suggestions for language improvements; to Shuo and Heng for the oriental perspective and for telling me that I use too many words.

Last but not least, I want to thank my friends and family for their outstanding support throughout the years. You know who you are.

Stavanger, Norway
April 2018

Krisztian Balog

Website

<http://eos-book.org>

This book is accompanied by the above website. The website provides a variety of supplementary material, corrections of mistakes, and related resources.

Contents

1	Introduction	1
1.1	What Is an Entity?	2
1.1.1	Named Entities vs. Concepts	3
1.1.2	Properties of Entities	4
1.1.3	Representing Properties of Entities	5
1.2	A Brief Historical Outlook	6
1.2.1	Information Retrieval	7
1.2.2	Databases	8
1.2.3	Natural Language Processing	9
1.2.4	Semantic Web	10
1.3	Entity-Oriented Search	11
1.3.1	A Bird’s-Eye View	11
1.3.2	Tasks and Challenges	14
1.3.3	Entity-Oriented vs. Semantic Search	15
1.3.4	Application Areas	16
1.4	About the Book	17
1.4.1	Focus	17
1.4.2	Audience and Prerequisites	17
1.4.3	Organization	18
1.4.4	Terminology and Notation	19
	References	20
2	Meet the Data	25
2.1	The Web	26
2.1.1	Datasets and Resources	27
2.2	Wikipedia	28
2.2.1	The Anatomy of a Wikipedia Article	29
2.2.2	Links	32
2.2.3	Special-Purpose Pages	33
2.2.4	Categories, Lists, and Navigation Templates	33
2.2.5	Resources	35

2.3	Knowledge Bases	36
2.3.1	A Knowledge Base Primer	37
2.3.2	DBpedia	40
2.3.3	YAGO	45
2.3.4	Freebase	46
2.3.5	Wikidata	47
2.3.6	The Web of Data	48
2.3.7	Standards and Resources	51
2.4	Summary	51
	References	52

Part I Entity Ranking

3	Term-Based Models for Entity Ranking	57
3.1	The Ad Hoc Entity Retrieval Task	58
3.2	Constructing Term-Based Entity Representations	59
3.2.1	Representations from Unstructured Document Corpora	61
3.2.2	Representations from Semi-structured Documents	67
3.2.3	Representations from Structured Knowledge Bases	69
3.3	Ranking Term-Based Entity Representations	74
3.3.1	Unstructured Retrieval Models	75
3.3.2	Fielded Retrieval Models	79
3.3.3	Learning-to-Rank	82
3.4	Ranking Entities Without Direct Representations	85
3.5	Evaluation	86
3.5.1	Evaluation Measures	86
3.5.2	Test Collections	88
3.6	Summary	94
3.7	Further Reading	94
	References	95
4	Semantically Enriched Models for Entity Ranking	101
4.1	Semantics Means Structure	103
4.2	Preserving Structure	104
4.2.1	Multi-Valued Predicates	105
4.2.2	References to Entities	107
4.3	Entity Types	111
4.3.1	Type Taxonomies and Challenges	111
4.3.2	Type-Aware Entity Ranking	113
4.3.3	Estimating Type-Based Similarity	113
4.4	Entity Relationships	116
4.4.1	Ad Hoc Entity Retrieval	116
4.4.2	List Search	118
4.4.3	Related Entity Finding	120

- 4.5 Similar Entity Search..... 124
 - 4.5.1 Pairwise Entity Similarity 126
 - 4.5.2 Collective Entity Similarity 130
- 4.6 Query-Independent Ranking..... 133
 - 4.6.1 Popularity 134
 - 4.6.2 Centrality..... 135
 - 4.6.3 Other Methods 138
- 4.7 Summary..... 139
- 4.8 Further Reading 139
- References..... 140

Part II Bridging Text and Structure

- 5 Entity Linking 147**
 - 5.1 From Named Entity Recognition Toward Entity Linking 148
 - 5.1.1 Named Entity Recognition 149
 - 5.1.2 Named Entity Disambiguation 150
 - 5.1.3 Entity Coreference Resolution 151
 - 5.2 The Entity Linking Task 152
 - 5.3 The Anatomy of an Entity Linking System 152
 - 5.4 Mention Detection..... 154
 - 5.4.1 Surface Form Dictionary Construction 155
 - 5.4.2 Filtering Mentions 156
 - 5.4.3 Overlapping Mentions..... 157
 - 5.5 Candidate Selection 157
 - 5.6 Disambiguation..... 159
 - 5.6.1 Features 159
 - 5.6.2 Approaches 164
 - 5.6.3 Pruning 172
 - 5.7 Entity Linking Systems 172
 - 5.8 Evaluation 174
 - 5.8.1 Evaluation Measures 174
 - 5.8.2 Test Collections 175
 - 5.8.3 Component-Based Evaluation 179
 - 5.9 Resources 180
 - 5.9.1 A Cross-Lingual Dictionary for English Wikipedia Concepts 180
 - 5.9.2 Freebase Annotations of the ClueWeb Corpora 180
 - 5.10 Summary..... 181
 - 5.11 Further Reading 182
 - References..... 183

6	Populating Knowledge Bases	189
6.1	Harvesting Knowledge from Text	191
6.1.1	Class-Instance Acquisition	192
6.1.2	Class-Attribute Acquisition	195
6.1.3	Relation Extraction	195
6.2	Entity-Centric Document Filtering	197
6.2.1	Overview	198
6.2.2	Mention Detection	199
6.2.3	Document Scoring	200
6.2.4	Features	203
6.2.5	Evaluation	207
6.3	Slot Filling	212
6.3.1	Approaches	213
6.3.2	Evaluation	215
6.4	Summary	215
6.5	Further Reading	216
	References	216
Part III Semantic Search		
7	Understanding Information Needs	225
7.1	Semantic Query Analysis	226
7.1.1	Query Classification	226
7.1.2	Query Annotation	228
7.1.3	Query Interpretation	231
7.2	Identifying Target Entity Types	232
7.2.1	Problem Definition	233
7.2.2	Unsupervised Approaches	234
7.2.3	Supervised Approach	236
7.2.4	Evaluation	236
7.3	Entity Linking in Queries	239
7.3.1	Entity Annotation Tasks	240
7.3.2	Pipeline Architecture for Interpretation Finding	242
7.3.3	Candidate Entity Ranking	243
7.3.4	Producing Interpretations	246
7.4	Query Templates	252
7.4.1	Concepts and Definitions	253
7.4.2	Template Discovery Methods	255
7.5	Summary	260
7.6	Further Reading	261
	References	261
8	Leveraging Entities in Document Retrieval	269
8.1	Mapping Queries to Entities	270
8.2	Leveraging Entities for Query Expansion	272
8.2.1	Document-Based Query Expansion	273
8.2.2	Entity-Centric Query Expansion	274

- 8.2.3 Unsupervised Term Selection 275
- 8.2.4 Supervised Term Selection 276
- 8.3 Projection-Based Methods 279
 - 8.3.1 Explicit Semantic Analysis 280
 - 8.3.2 Latent Entity Space Model 282
 - 8.3.3 EsdRank 283
- 8.4 Entity-Based Representations 285
 - 8.4.1 Entity-Based Document Language Models 285
 - 8.4.2 Bag-of-Entities Representation 287
- 8.5 Practical Considerations 292
- 8.6 Resources and Test Collections 292
- 8.7 Summary 293
- 8.8 Further Reading 293
- References 294
- 9 Utilizing Entities for an Enhanced Search Experience 299**
 - 9.1 Query Assistance 299
 - 9.1.1 Query Auto-completion 300
 - 9.1.2 Query Recommendations 302
 - 9.1.3 Query Building Interfaces 310
 - 9.2 Entity Cards 312
 - 9.2.1 The Anatomy of an Entity Card 313
 - 9.2.2 Factual Entity Summaries 314
 - 9.3 Entity Recommendations 319
 - 9.3.1 Recommendations Given an Entity 320
 - 9.3.2 Personalized Recommendations 322
 - 9.3.3 Contextual Recommendations 325
 - 9.3.4 Explaining Recommendations 327
 - 9.4 Summary 331
 - 9.5 Further Reading 332
 - References 332
- 10 Conclusions and Future Directions 337**
 - 10.1 Summary of Progress 338
 - 10.1.1 Data 338
 - 10.1.2 Retrieval Methods 338
 - 10.1.3 Understanding and Interacting with Users 339
 - 10.2 A Peek into the Future 340
 - 10.3 Future Research Directions 343
 - 10.3.1 Understanding and Interacting with Users 344
 - 10.3.2 Complex Information Needs and Task Completion 345
 - 10.3.3 Data and Knowledge 346
 - 10.4 Concluding Remarks 346
 - References 347
- Index 349**

Acronyms

EF	Entity frequency
EL	Entity linking
ELQ	Entity linking in query
ER	Entity retrieval
IEF	Inverse entity frequency
INEX	Initiative for the Evaluation of XML Retrieval
IR	Information retrieval
KB	Knowledge base
KG	Knowledge graph
KR	Knowledge repository
LM	Language models
LTR	Learning-to-rank
NLP	Natural language processing
SDM	Sequential dependence model
SERP	Search engine result page
SPO	Subject-predicate-object (triple)
TREC	Text Retrieval Conference

Notation

Throughout this book, unless stated otherwise, the notation used is as follows:

Symbol	Meaning
$c(x)$	Total count of x
$c(x; y)$	Count of x in the context of y
$c(x, y; z)$	Number of times x and y co-occur in the context of z
d	Document ($d \in \mathcal{D}$)
\mathcal{D}	Document collection
$\mathcal{D}_q(k)$	Top- k ranked documents for query q
e	Entity ($e \in \mathcal{E}$)
\mathcal{E}	Entity catalog (set of all entities)
$\mathcal{E}_q(k)$	Top- k ranked entities for query q
\mathcal{K}	Knowledge base (set of SPO triples)
\mathcal{L}_e	Set of links of an entity e
l_x	Representation length of x ($l_x = \sum_{t \in \mathcal{V}} c(t; x)$)
q	Query
t	Term (string token, $t \in \mathcal{V}$)
\mathcal{T}_e	Types of entity e ($\mathcal{T}_e \subset \mathcal{T}$)
\mathcal{T}	Type taxonomy
\mathcal{V}	Vocabulary of terms
$ X $	Cardinality of set X
Z	Normalization factor
$\mathbb{1}(x)$	Binary indicator function (returns 1 if x is true, otherwise 0)

Chapter 1

Introduction



Search engines have become part of our daily lives. We use Google (Bing, Yandex, Baidu, etc.) as the main gateway to find information on the Web. With a certain type of content in mind, we may search directly on a particular site or service, e.g., on Facebook or LinkedIn for people, organizations, and events; on Amazon or eBay for products; or on YouTube or Spotify for music. Even on our smartphones, we are increasingly reliant on search functionality to find contacts, email, notes, calendar entries, apps, etc. We have grown accustomed to expect a search box somewhere near the top of the screen, and we have also increased our expectations of the quality and speed of the responses to our searches.

On the highest level of abstraction, the field of information retrieval (IR) is concerned with developing technology for matching *information needs* with *information objects*. What we put in the search box, i.e., the *query*, is an expression of our information need. It may range from a few simple keywords (e.g., “*Bond girls*”) to a proper natural language question (e.g., “*What are good digital cameras under \$300?*”). The search engine then responds with a ranked list of items, i.e., information objects. Traditionally, these items were documents. In fact, IR has been seen as synonymous with document retrieval by many. The past decade, however, has seen an enormous development in search technology. As regular users, we have witnessed first-hand the transition of search engines into “answering engines.” Today’s contemporary web search engines return rich search result pages, which include direct displays of entities, facts, and other structured results instead of merely a list of documents (“ten blue links”), as illustrated in Fig. 1.1. A primary enabling component behind these advanced search services is the availability of large-scale structured knowledge repositories (called *knowledge bases*), which organize information around specific things or objects (which we will be referring to as *entities*). The objective of this book is to give a detailed account of the developments of a decade of IR research that have enabled us to search for “things, not strings.”

Thank You for previewing this eBook

You can read the full version of this eBook in different formats:

- HTML (Free /Available to everyone)
- PDF / TXT (Available to V.I.P. members. Free Standard members can access up to 5 PDF/TXT eBooks per month each month)
- Epub & Mobipocket (Exclusive to V.I.P. members)

To download this full book, simply select the format you desire below

