The Information Retrieval Series

Krisztian Balog

Entity-Oriented Search

Springer Open

The Information Retrieval Series Volume 39

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

Krisztian Balog

Entity-Oriented Search

Springer Open

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

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

[1] Note 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 Krisztian Balog
April 2018

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

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 transitioning 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."

K. Balog, *Entity-Oriented Search*, The Information Retrieval Series 39,
https://doi.org/10.1007/978-3-319-93935-3_1