Introduction to Information Retrieval

Introduction to Information Retrieval is the first textbook with a coherent treatment of classical and web information retrieval, including web search and the related areas of text classification and text clustering. Written from a computer science perspective, it gives an up-to-date treatment of all aspects of the design and implementation of systems for gathering, indexing, and searching documents and of methods for evaluating systems, along with an introduction to the use of machine learning methods on text collections.

Designed as the primary text for a graduate or advanced undergraduate course in information retrieval, the book will also interest researchers and professionals. A complete set of lecture slides and exercises that accompany the book are available on the web.

Christopher D. Manning is Associate Professor of Computer Science and Linguistics at Stanford University.

Prabhakar Raghavan is Head of Yahoo! Research and a Consulting Professor of Computer Science at Stanford University.

Hinrich Schütze is Chair of Theoretical Computational Linguistics at the Institute for Natural Language Processing, University of Stuttgart.

Introduction to Information Retrieval

Christopher D. Manning
Stanford University

Prabhakar Raghavan
Yahoo! Research

Hinrich Schütze
University of Stuttgart

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore, São Paulo, Delhi

Cambridge University Press
32 Avenue of the Americas, New York, NY 10013-2473, USA

www.cambridge.org
Information on this title: www.cambridge.org/9780521865715

First published 2008

Printed in the United States of America

A catalog record for this publication is available from the British Library.

Library of Congress Cataloging in Publication data

Manning, Christopher D.
Introduction to information retrieval / Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze.
p. cm.
Includes bibliographical references and index.
ISBN 978-0-521-86571-5 (hardback)
1. Text processing (Computer science) 2. Information retrieval. 3. Document clustering. 4. Semantic Web. I. Raghavan, Prabhakar. II. Schütze, Hinrich. III. Title.
QA76.9.T48M26 2008
025.04 – dc22 2008001257

ISBN 978-0-521-86571-5 hardback

Contents

Table of Notation

Symbol	Page	Meaning
γ	90	γ code
γ	237	Classification or clustering function: $\gamma(d)$ is d's class or cluster
Γ	237	Supervised learning method in Chapters 13 and 14: $\Gamma(\mathbb{D})$ is the classification function γ learned from training set $\mathbb{D}$
λ	370	Eigenvalue
$\vec{\mu}(.)$	269	Centroid of a class (in Rocchio classification) or a cluster (in K-means and centroid clustering)
Φ	105	Training example
σ	374	Singular value
$\Theta(\cdot)$	10	A tight bound on the complexity of an algorithm
ω, ω_k	328	Cluster in clustering
Ω	328	Clustering or set of clusters $\{\omega_1, \ldots, \omega_K\}$
$\arg\max_x f(x)$	164	The value of x for which f reaches its maximum
$\arg\min_x f(x)$	164	The value of x for which f reaches its minimum
c, c_j	237	Class or category in classification
cf_t	82	The collection frequency of term t (the total number of times the term appears in the document collection)
$\mathbb{C}$	237	Set $\{c_1, \ldots, c_J\}$ of all classes
C	248	A random variable that takes as values members of $\mathbb{C}$
C	369	Term–document matrix
d	4	Index of the dth document in the collection D
d	65	A document
$\vec{d}, \vec{q}$	163	Document vector, query vector
D	326	Set $\{d_1, \ldots, d_N\}$ of all documents
D_c	269	Set of documents that is in class c
$\mathbb{D}$	237	Set $\{\langle d_1, c_1\rangle, \ldots, \langle d_N, c_N\rangle\}$ of all labeled documents in Chapters 13–15

df_t	108	The document frequency of term t (the total number of documents in the collection the term appears in)
H	91	Entropy
H_M	93	Mth harmonic number
$I(X;Y)$	252	Mutual information of random variables X and Y
idf_t	108	Inverse document frequency of term t
J	237	Number of classes
k	267	Top k items from a set, e.g., k nearest neighbors in kNN, top k retrieved documents, top k selected features from the vocabulary V
k	50	Sequence of k characters
K	326	Number of clusters
L_d	214	Length of document d (in tokens)
L_a	242	Length of the test document (or application document) in tokens
L_ave	64	Average length of a document (in tokens)
M	4	Size of the vocabulary ($\|V\|$)
M_a	242	Size of the vocabulary of the test document (or application document)
M_ave	71	Average size of the vocabulary in a document in the collection
M_d	218	Language model for document d
N	4	Number of documents in the retrieval or training collection
N_c	240	Number of documents in class c
$N(\omega)$	275	Number of times the event ω occurred
$O(\cdot)$	10	A bound on the complexity of an algorithm
$O(\cdot)$	203	The odds of an event
P	142	Precision
$P(\cdot)$	202	Probability
P	425	Transition probability matrix
q	55	A query
R	143	Recall
s_i	53	A string
s_i	103	Boolean values for zone scoring
$\mathrm{sim}(d_1, d_2)$	111	Similarity score for documents d_1, d_2
T	40	Total number of tokens in the document collection
T_{ct}	240	Number of occurrences of word t in documents of class c
t	4	Index of the tth term in the vocabulary V
t	56	A term in the vocabulary
$\mathrm{tf}_{t,d}$	107	The term frequency of term t in document d (the total number of occurrences of t in d)

U_t	246	Random variable taking values 0 (term t is present) and 1 (t is not present)
V	190	Vocabulary of terms $\{t_1, \ldots, t_M\}$ in a collection (a.k.a. the lexicon)
$\vec{v}(d)$	111	Length-normalized document vector
$\vec{V}(d)$	110	Vector of document d, not length normalized
$\text{wf}_{t,d}$	115	Weight of term t in document d
w	103	A weight, for example, for zones or terms
$\vec{w}^{\mathrm{T}}\vec{x} = b$	269	Hyperplane; $\vec{w}$ is the normal vector of the hyperplane and w_i component i of $\vec{w}$
$\vec{x}$	204	Term incidence vector $\vec{x} = (x_1, \ldots, x_M)$; more generally: document feature representation
X	246	Random variable taking values in V, the vocabulary (e.g., at a given position k in a document)
$\mathbb{X}$	237	Document space in text classification
$\lvert A\rvert$	56	Set cardinality: the number of members of set A
$\lvert S\rvert$	370	Determinant of the square matrix S
$\lvert s_i\rvert$	53	Length in characters of string s_i
$\lvert\vec{x}\rvert$	128	Length of vector $\vec{x}$
$\lvert\vec{x} - \vec{y}\rvert$	121	Euclidean distance of $\vec{x}$ and $\vec{y}$ (which is the length of $(\vec{x} - \vec{y})$)

Preface

As recently as the 1990s, studies showed that most people preferred getting information from other people rather than from information retrieval (IR) systems. Of course, in that time period, most people also used human travel agents to book their travel. However, during the last decade, relentless optimization of information retrieval effectiveness has driven web search engines to new quality levels at which most people are satisfied most of the time, and web search has become a standard and often preferred source of information finding. For example, the 2004 Pew Internet Survey (Fallows 2004) found that "92% of Internet users say the Internet is a good place to go for getting everyday information." To the surprise of many, the field of information retrieval has moved from being a primarily academic discipline to being the basis underlying most people's preferred means of information access. This book presents the scientific underpinnings of this field, at a level accessible to graduate students as well as advanced undergraduates.

Information retrieval did not begin with the Web. In response to various challenges of providing information access, the field of IR evolved to give principled approaches to searching various forms of content. The field began with scientific publications and library records but soon spread to other forms of content, particularly those of information professionals, such as journalists, lawyers, and doctors. Much of the scientific research on IR has occurred in these contexts, and much of the continued practice of IR deals with providing access to unstructured information in various corporate and governmental domains, and this work forms much of the foundation of our book.

Nevertheless, in recent years, a principal driver of innovation has been the World Wide Web, unleashing publication at the scale of tens of millions of content creators. This explosion of published information would be moot if the information could not be found, annotated, and analyzed so that each user can quickly find information that is both relevant and comprehensive for their needs. By the late 1990s, many people felt that continuing to index the whole Web would rapidly become impossible, due to the Web's

exponential growth in size. But major scientific innovations, superb engineering, the rapidly declining price of computer hardware, and the rise of a commercial underpinning for web search have all conspired to power today's major search engines, which are able to provide high-quality results within subsecond response times for hundreds of millions of searches a day over billions of web pages.

Book organization and course development

This book is the result of a series of courses we have taught at Stanford University and at the University of Stuttgart, in a range of durations including a single quarter, one semester, and two quarters. These courses were aimed at early stage graduate students in computer science, but we have also had enrollment from upper-class computer science undergraduates, as well as students from law, medical informatics, statistics, linguistics, and various engineering disciplines. The key design principle for this book, therefore, was to cover what we believe to be important in a one-term graduate course on IR. An additional principle is to build each chapter around material that we believe can be covered in a single lecture of 75 to 90 minutes.

The first eight chapters of the book are devoted to the basics of information retrieval and in particular the heart of search engines; we consider this material to be core to any course on information retrieval. Chapter 1 introduces inverted indexes and shows how simple Boolean queries can be processed using such indexes. Chapter 2 builds on this introduction by detailing the manner in which documents are preprocessed before indexing and by discussing how inverted indexes are augmented in various ways for functionality and speed. Chapter 3 discusses search structures for dictionaries and how to process queries that have spelling errors and other imprecise matches to the vocabulary in the document collection being searched. Chapter 4 describes a number of algorithms for constructing the inverted index from a text collection with particular attention to highly scalable and distributed algorithms that can be applied to very large collections. Chapter 5 covers techniques for compressing dictionaries and inverted indexes. These techniques are critical for achieving subsecond response times to user queries in large search engines. The indexes and queries considered in Chapters 1 through 5 only deal with *Boolean retrieval*, in which a document either matches a query or does not. A desire to measure the *extent* to which a document matches a query, or the score of a document for a query, motivates the development of term weighting and the computation of scores in Chapters 6 and 7, leading to the idea of a list of documents that are rank-ordered for a query. Chapter 8 focuses on the evaluation of an information retrieval system based on the relevance of the documents it retrieves, allowing us to compare the relative

performances of different systems on benchmark document collections and queries.

Chapters 9 through 21 build on the foundation of the first eight chapters to cover a variety of more advanced topics. Chapter 9 discusses methods by which retrieval can be enhanced through the use of techniques like relevance feedback and query expansion, which aim at increasing the likelihood of retrieving relevant documents. Chapter 10 considers IR from documents that are structured with markup languages like XML and HTML. We treat structured retrieval by reducing it to the vector space scoring methods developed in Chapter 6. Chapters 11 and 12 invoke probability theory to compute scores for documents on queries. Chapter 11 develops traditional probabilistic IR, which provides a framework for computing the probability of relevance of a document, given a set of query terms. This probability may then be used as a score in ranking. Chapter 12 illustrates an alternative, wherein, for each document in a collection, we build a language model from which one can estimate a probability that the language model generates a given query. This probability is another quantity with which we can rank-order documents.

Chapters 13 through 18 give a treatment of various forms of machine learning and numerical methods in information retrieval. Chapters 13 through 15 treat the problem of classifying documents into a set of known categories, given a set of documents along with the classes they belong to. Chapter 13 motivates statistical classification as one of the key technologies needed for a successful search engine; introduces Naive Bayes, a conceptually simple and efficient text classification method; and outlines the standard methodology for evaluating text classifiers. Chapter 14 employs the vector space model from Chapter 6 and introduces two classification methods, Rocchio and k nearest neighbor (kNN), that operate on document vectors. It also presents the bias-variance tradeoff as an important characterization of learning problems that provides criteria for selecting an appropriate method for a text classification problem. Chapter 15 introduces support vector machines, which many researchers currently view as the most effective text classification method. We also develop connections in this chapter between the problem of classification and seemingly disparate topics such as the induction of scoring functions from a set of training examples.

Chapters 16, 17, and 18 consider the problem of inducing clusters of related documents from a collection. In Chapter 16, we first give an overview of a number of important applications of clustering in IR. We then describe two flat clustering algorithms: the K-means algorithm, an efficient and widely used document clustering method, and the expectation-maximization algorithm, which is computationally more expensive, but also more flexible. Chapter 17 motivates the need for hierarchically structured clusterings (instead of flat clusterings) in many applications in IR and introduces a number of clustering algorithms that produce a hierarchy of clusters. The chapter

also addresses the difficult problem of automatically computing labels for clusters. Chapter 18 develops methods from linear algebra that constitute an extension of clustering and also offer intriguing prospects for algebraic methods in IR, which have been pursued in the approach of latent semantic indexing.

Chapters 19 through 21 treat the problem of web search. We give in Chapter 19 a summary of the basic challenges in web search, together with a set of techniques that are pervasive in web information retrieval. Next, Chapter 20 describes the architecture and requirements of a basic web crawler. Finally, Chapter 21 considers the power of link analysis in web search, using in the process several methods from linear algebra and advanced probability theory.

This book is not comprehensive in covering all topics related to IR. We have put aside a number of topics, which we deemed outside the scope of what we wished to cover in an introduction to IR class. Nevertheless, for people interested in these topics, we provide the following pointers to mainly textbook coverage:

Cross-language IR Grossman and Frieder 2004, ch. 4, and Oard and Dorr 1996.

Image and multimedia IR Grossman and Frieder 2004, ch. 4; Baeza-Yates and Ribeiro-Neto 1999, ch. 6; Baeza-Yates and Ribeiro-Neto 1999, ch. 11; Baeza-Yates and Ribeiro-Neto 1999, ch. 12; del Bimbo 1999; Lew 2001; and Smeulders et al. 2000.

Speech retrieval Coden et al. 2002.

Music retrieval Downie 2006 and http://www.ismir.net/.

User interfaces for IR Baeza-Yates and Ribeiro-Neto 1999, ch. 10.

Parallel and peer-to-peer IR Grossman and Frieder 2004, ch. 7; Baeza-Yates and Ribeiro-Neto 1999, ch. 9; and Aberer 2001.

Digital libraries Baeza-Yates and Ribeiro-Neto 1999, ch. 15, and Lesk 2004.

Information science perspective Korfhage 1997; Meadow et al. 1999; and Ingwersen and Järvelin 2005.

Logic-based approaches to IR van Rijsbergen 1989.

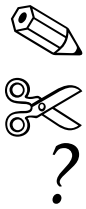
Natural language processing techniques Manning and Schütze 1999; Jurafsky and Martin 2008; and Lewis and Jones 1996.

Prerequisites

Introductory courses in data structures and algorithms, in linear algebra, and in probability theory suffice as prerequisites for all twenty-one chapters. We now give more detail for the benefit of readers and instructors who wish to tailor their reading to some of the chapters.

Chapters 1 through 5 assume as prerequisite a basic course in algorithms and data structures. Chapters 6 and 7 require, in addition, a knowledge of basic linear algebra, including vectors and dot products. No additional prerequisites are assumed until Chapter 11, for which a basic course in probability theory is required; Section 11.1 gives a quick review of the concepts necessary in Chapters 11, 12, and 13. Chapter 15 assumes that the reader is familiar with the notion of nonlinear optimization, although the chapter may be read without detailed knowledge of algorithms for nonlinear optimization. Chapter 18 demands a first course in linear algebra, including familiarity with the notions of matrix rank and eigenvectors; a brief review is given in Section 18.1. The knowledge of eigenvalues and eigenvectors is also necessary in Chapter 21.

Book layout



Worked examples in the text appear with a pencil sign next to them in the left margin. Advanced or difficult material appears in sections or subsections indicated with scissors in the margin. Exercises are marked in the margin with a question mark. The level of difficulty of exercises is indicated as easy [⋆], medium [⋆⋆], or difficult [⋆ ⋆ ⋆].

Acknowledgments

The authors thank Cambridge University Press for allowing us to make the draft book available online, which facilitated much of the feedback we have received while writing the book. We also thank Lauren Cowles, who has been an outstanding editor, providing several rounds of comments on each chapter; on matters of style, organization, and coverage; as well as detailed comments on the subject matter of the book. To the extent that we have achieved our goals in writing this book, she deserves an important part of the credit.

We are very grateful to the many people who have given us comments, suggestions, and corrections based on draft versions of this book. We thank for providing various corrections and comments: Cheryl Aasheim, Josh Attenberg, Luc Bélanger, Tom Breuel, Daniel Burckhardt, Georg Buscher, Fazli Can, Dinquan Chen, Ernest Davis, Pedro Domingos, Rodrigo Panchiniak Fernandes, Paolo Ferragina, Norbert Fuhr, Vignesh Ganapathy, Elmer Garduno, Xiubo Geng, David Gondek, Sergio Govoni, Corinna Habets, Ben Handy, Donna Harman, Benjamin Haskell, Thomas Hühn, Deepak Jain, Ralf Jankowitsch, Dinakar Jayarajan, Vinay Kakade, Mei Kobayashi, Wessel Kraaij, Rick Lafleur, Florian Laws, Hang Li, David Mann, Ennio Masi, Frank McCown, Paul McNamee, Sven Meyer zu Eissen, Alexander Murzaku, Gonzalo Navarro, Scott Olsson, Daniel Paiva, Tao Qin, Megha Raghavan,

Ghulam Raza, Michal Rosen-Zvi, Klaus Rothenhäusler, Kenyu L. Runner, Alexander Salamanca, Grigory Sapunov, Tobias Scheffer, Nico Schlaefer, Evgeny Shadchnev, Ian Soboroff, Benno Stein, Marcin Sydow, Andrew Turner, Jason Utt, Huey Vo, Travis Wade, Mike Walsh, Changliang Wang, Renjing Wang, and Thomas Zeume.

Many people gave us detailed feedback on individual chapters, either at our request or through their own initiative. For this, we're particularly grateful to James Allan, Omar Alonso, Ismail Sengor Altingovde, Vo Ngoc Anh, Roi Blanco, Eric Breck, Eric Brown, Mark Carman, Carlos Castillo, Junghoo Cho, Aron Culotta, Doug Cutting, Meghana Deodhar, Susan Dumais, Johannes Fürnkranz, Andreas Heß, Djoerd Hiemstra, David Hull, Thorsten Joachims, Siddharth Jonathan J. B., Jaap Kamps, Mounia Lalmas, Amy Langville, Nicholas Lester, Dave Lewis, Stephen Liu, Daniel Lowd, Yosi Mass, Jeff Michels, Alessandro Moschitti, Amir Najmi, Marc Najork, Giorgio Maria Di Nunzio, Paul Ogilvie, Priyank Patel, Jan Pedersen, Kathryn Pedings, Vassilis Plachouras, Daniel Ramage, Stefan Riezler, Michael Schiehlen, Helmut Schmid, Falk Nicolas Scholer, Sabine Schulte im Walde, Fabrizio Sebastiani, Sarabjeet Singh, Alexander Strehl, John Tait, Shivakumar Vaithyanathan, Ellen Voorhees, Gerhard Weikum, Dawid Weiss, Yiming Yang, Yisong Yue, Jian Zhang, and Justin Zobel.

And finally there were a few reviewers who absolutely stood out in terms of the quality and quantity of comments that they provided. We thank them for their significant impact on the content and structure of the book. We express our gratitude to Pavel Berkhin, Stefan Büttcher, Jamie Callan, Byron Dom, Torsten Suel, and Andrew Trotman.

Parts of the initial drafts of Chapters 13, 14, and 15 were based on slides that were generously provided by Ray Mooney. Although the material has gone through extensive revisions, we gratefully acknowledge Ray's contribution to the three chapters in general and to the description of the time complexities of text classification algorithms in particular.

The above is unfortunately an incomplete list; we are still in the process of incorporating feedback we have received. And, like all opinionated authors, we did not always heed the advice that was so freely given. The published versions of the chapters remain solely the responsibility of the authors.

The authors thank Stanford University and the University of Stuttgart for providing a stimulating academic environment for discussing ideas and the opportunity to teach courses from which this book arose and in which its contents were refined. CM thanks his family for the many hours they've let him spend working on this book and hopes he'll have a bit more free time on weekends next year. PR thanks his family for their patient support through the writing of this book and is also grateful to Yahoo! Inc. for providing a fertile environment in which to work on this book. HS would like to thank his parents, family, and friends for their support while writing this book.

Web and contact information

This book has a companion website at http://informationretrieval.org. As well as links to some more general resources, it is our intention to maintain on this website a set of slides for each chapter that may be used for the corresponding lecture. We gladly welcome further feedback, corrections, and suggestions on the book, which may be sent to all the authors at informationretrieval@yahoogroups.com.