Modern Information Retrieval

the concepts and technology behind search

Second edition

Ricardo Baeza-Yates

Berthier Ribeiro-Neto

Modern Information Retrieval

We work with leading authors to develop the strongest educational materials in information retrieval, bringing cutting-edge thinking and best learning practice to a global market.

Under a range of well-known imprints, including Addison Wesley, we craft high quality print and electronic publications which help readers to understand and apply their content, whether studying or at work.

To find out more about the complete range of our publishing, please visit us on the World Wide Web at:
www.pearsoned.co.uk

Modern Information Retrieval

the concepts and technology behind search

Second edition

Ricardo Baeza-Yates
Berthier Ribeiro-Neto

Addison Wesley
is an imprint of

Harlow, England • London • New York • Boston • San Francisco • Toronto • Sydney • Singapore • Hong Kong
Tokyo • Seoul • Taipei • New Delhi • Cape Town • Madrid • Mexico City • Amsterdam • Munich • Paris • Milan

Pearson Education Limited
Edinburgh Gate
Harlow
Essex CM20 2JE
England

and Associated Companies throughout the world

Visit us on the World Wide Web at:
www.pearsoned.co.uk

First published 1999
Second edition published 2011

ISBN: 978-0-321-41691-9

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Cataloging-in-Publication Data
A catalog record for this book is available from the Library of Congress

10 9 8 7 6 5 4 3 2
13 12 11

Typeset in CMR10 10/12 by 73
Printed and bound in Great Britain by Henry Ling Ltd at the Dorset Press, Dorchester, Dorset

About the cover: the word cloud was produced using the frequency of keywords in the text content of the book, using wordle.net

Contents

Preface to the Second Edition

Since the first edition of this book, much has happened in the Information Retrieval (IR) arena, particularly with regard to the Web. For once, the gargantuan volume of information on the Web has transformed the search engines into key tools for seeking and finding information of interest. Further, since the search engines are fundamentally IR systems at their very core, they have become the main proof-of-concept of the application of IR technologies to vast document collections with huge query traffic.

We closely followed this evolution by starting search engines in Brazil and Chile, just a few months after the first edition appeared. Later on, we joined the two major search engine companies, Google and Yahoo!, getting even closer to all the action. Hence, this second edition of *Modern Information Retrieval* reflects not only the changes in the IR field, but also our own experience in research, development, and execution of IR technology, particularly when applied to the Web.

The first edition of *Modern Information Retrieval* was not a book written in standard fashion, given that we asked contributors to write chapters in areas in which we felt we did not have enough expertise. So, in some sense, we preceded the Web 2.0 trend of development in collaboration with a team. We aimed at a well integrated book by carefully coordinating and supervising all the writing. To a certain extent, our efforts worked well. Indeed, the first edition of the book sold very well and became the IR best seller, having been reprinted many times. The book has been adopted by hundreds of universities and schools. It has also been translated first to Korean and then to Chinese, with an special non-expensive edition having been printed in India. Hence, just a couple of years after the first edition was printed, we started talking about a second edition. The idea did not materialize until 2004 when we submitted a proposal to the publisher, which was approved. We eventually started working on the second edition by November 2005, more than five years ago. Today, we have finally finished!

In this second edition of *Modern Information Retrieval* we have followed the same methodology, as it clearly worked with the first edition. Nonetheless, in this case we are authors or co-authors of more chapters and we have taken a stronger hand in shaping the content of the contributed chapters. As a consequence, we have had to change completely many chapters and add many new ones. As a result, 60–70% of this second edition is made of new material, which mostly differs from the first edition

in the following aspects:

1. A complete reorganization of the content of the first chapters.
2. New chapters on text classification, Web crawling, structured text retrieval, and enterprise search plus a new appendix on open source search engines.
3. Fully rewritten chapters on user interfaces, multimedia retrieval, and digital libraries.
4. Expanded chapters to include new and important developments such as language models, new evaluation measures, characteristics of queries, cluster-based and distributed IR, learning to rank, search engines interfaces, click-through data, and personalization.
5. An improved Web site, see www.mir2ed.org, aimed at becoming a reference IR teaching resource, which includes a full set of slides for all chapters in the book and recommended lists of exercises.

The final outcome is a book that is almost twice as long and that contains more than twice the number of references of the first edition. In summary, if you liked the first edition of *Modern Information Retrieval*, we hope you will like this second edition even more. And, in case you did not like the first edition, we hope that this time you will change your mind.

Ricardo Baeza-Yates, Barcelona, Spain
Berthier Ribeiro-Neto, Belo Horizonte, Brazil
December, 2010

Preface to the First Edition

Information retrieval (IR) has changed considerably in the last years with the expansion of the Web (World Wide Web) and the advent of modern and inexpensive graphical user interfaces and mass storage devices. As a result, traditional IR textbooks have become quite out-of-date which has led to the introduction of new IR books recently. Nevertheless, we believe that there is still great need of a book that approaches the field in a rigorous and complete way from a computer-science perspective (in opposition to a user-centered perspective). This book is an effort to partially fulfill this gap and should be useful for a first course on information retrieval as well as for a graduate course on the topic.

The book is composed of two portions which complement and balance each other. The core portion includes 9 chapters authored or coauthored by the designers of the book. The second portion, which is fully integrated with the first, is formed by 6 state-of-the-art chapters written by leading researchers in their fields. A same notation and glossary are employed in all the chapters. Thus, despite the fact that several people contributed to the text, this book is really much more a textbook than an edited collection of chapters written by separate authors. Further, contrary to a collection of chapters, the contents and organization of this book have been carefully designed by the main authors to present a cohesive view of all the important aspects of modern information retrieval.

From IR models to indexing text, from IR visual tools and interfaces to the Web, from IR multimedia to digital libraries, the book provides both broadness of coverage and richness of details. It is our hope that, given the now clear relevance and significance of information retrieval to modern society, the book will contribute to further disseminate the study of the discipline at information science, computer science, and library science departments throughout the world.

Ricardo Baeza-Yates, Santiago, Chile
Berthier Ribeiro-Neto, Belo Horizonte, Brazil
October, 1998

Authors' Acknowledgements to the Second Edition

We would like to sincerely thank the various people who, over a period of several years, provided us with useful and helpful comments, reviews, and suggestions. The improvements in the book content and in the organization of the material are largely due to them. Without their help, this second edition would not be of the same quality. Any errors that remain, hopefully only few, are entirely our responsibility.

First, we would like to thank all the chapter contributors, for their dedication and interest. To Eric Brown, Carlos Castillo, Marcos Gonçalves, David Hawking, Marti Hearst, Mounia Lalmas, Yoelle Maarek, Christian Middleton, Gonzalo Navarro, Dulce Ponceleón, Edie Rasmussen, Malcolm Slaney, and Nivio Ziviani, whose contributions reflect expertise we certainly have not fully mastered ourselves.

Second, to all the people who directly or indirectly contributed or influenced the new content of this second edition. We have to thank Omar Alonso (who pointed out that we were leaving out the important trend of crowdsourcing), Paolo Boldi (Web graph compression), Pavel Calado (text classification), Marco Cristo (whose comments on the text classification chapter led to an overall organization of the material), Christos Faloutsos (multidimensional indexing), Winston Hsu (multimedia), Flavio Junqueira (distributed retrieval), Edleno Moura (retrieval evaluation), Vanessa Murdock (query difficulty), Martin Porter (his stemming algorithm), Mark Sanderson (whose sharp comments led to great improvements on the retrieval evaluation chapter), Fabrizio Silvestri (URL ordering), and Gleb Skobeltsyn (P2P IR). Further, we also acknowledge the contributions of various graduate students of Marcos Gonçalves at the Federal University of Minas Gerais, Brazil, who reviewed and wrote extensive comments on the text classification chapter.

Third, we need to thank all the people who sent us errata for the first edition, comments for improvements and also comments on drafts of the second edition. In the case of errata we mention the first people who detected a mistake, as otherwise the list would be too long. They are, with the risk of omitting someone, Omar Alonso, Jose Hilario Canos, Berkant Barla Cambazoglu, Ernie Davis, Anne Diekema, Bill Dimm, Joaquim Gabarro, Jamie Geddes, Eduardo Graells, Kyoung-Soo Han, Claudia Hauff, Shoujie He, Ben Houston, Puay-Leng Lee, Songwook Lee, Shian-Hua Lin,

Mildrid Ljosland, Chang-Tien Lu, Mari Carmen Marcos, Peter Mika, Vanessa Murdock, Joanna Plattner, Luz Rello, Hee-Cheol Seo, Ben Shneiderman, Helge Grenager Solheim, Ellen Spertus, Markus Stocker, Kazunari Sugiyama, Satoru Takabayashi, Juha Takkinen, Luong Minh Thang, Yannis Tzitzikas, Fredrik Wallenberg, Theo van der Weide, John Westbrook, Judith Winter, Sui Xi, Peng Yong, Hugo Zaragoza, and Yonghui Zhang.

Fourth, special thanks to David Fernandes who made the teaching slides that can be found on the book Web site and patiently pointed out many small errors and inconsistencies throughout the book. Also, we need to mention the implicit support of our employers, Yahoo! and Google, in the always difficult task of writing a book.

Fifth, we have to thank our editors at Pearson Education. We started with Kate Brewin, then Simon Plumtree, next Owen Knight, and finally Rufus Curnow, who supported us during the most important part of the publishing process. During this process we had the help of Anita Atkinson as our desk editor and Jenny Oates as proof reader.

Finally, and most important, to Helena, Rosa, and our children, who once more put up with a string of trips abroad, lost weekends, and odd working hours. During these last four years, their recurrent question was: when will you finish the book?!

Authors' Acknowledgements to the First Edition

We would like to sincerely thank the various people who, throughout the several months in which this endeavor lasted, provided us with useful and helpful assistance. Without their care and consideration, this book would likely not have matured.

First, we would like to thank all the chapter contributors, for their dedication and interest. To Elisa Bertino, Eric Brown, Barbara Catania, Christos Faloutsos, Elena Ferrari, Ed Fox, Marti Hearst, Gonzalo Navarro, Edie Rasmussen, Ohm Sornil, and Nivio Ziviani, whose contributions reflect expertise we certainly have not fully mastered ourselves. And for all their patience throughout an editing and cross-reviewing process which constitutes a rather difficult balancing act.

Second, we would like to thank all the people who expressed interest in publishing this book, in particular, Scott Delman and Doug Sery.

Third, we would like to commend the interest, encouragement, and great job done by Addison Wesley Longman throughout the overall process, represented by Keith Mansfield, Karen Sutherland, Bridget Allen, David Harrison, Sheila Chatten, Helen Hodge, and Lisa Talbot. The reviewers they contacted read an early (and rather preliminary) proposal of this book and provided us with nice feedback and invaluable insights. The chapter on Parallel and Distributed IR was moved from the part on Applications of IR (where it did not fit well) to the part on Text IR due to the objective arguments of an unknown referee. A separate chapter on Retrieval Evaluation was only included after another zealous referee strongly made the case for the importance of this subject.

Fourth, we would like to thank all the people who discussed this project with us. Doug Oard provided us with an early critique of the proposal. Gary Marchionini was an earlier supporter and provided us with useful contacts during the process. Bruce Croft encouraged our effort since the beginning. Alberto Mendelzon provided us with an initial proposal and a compilation of references for the chapter on searching the Web. Ed Fox found time in a rather busy schedule to provide us with an insightful review of the introduction (which resulted in a great improvement) and a thorough review of the chapter on Modeling. Marti Hearst expressed interest in our proposal early on, provided assistance throughout the editing process, and has been an enthusiastic supporter and partner.

Fifth, we thank the support of our institutions, the Departments of Computer Science of the University of Chile and of the Federal University of Minas Gerais, as well as the funding provided by national research agencies (CNPq in Brazil and CONICYT in Chile) and international collaboration projects, in particular CYTED project VII.13 AMYRI (Environment for Information Managing and Retrieval in the World Wide Web) and Finep project SIAM (Information Systems for Mobile Computers).

Most important, to Helena, Rosa, and our children, who put up with a string of trips abroad, lost weekends, and odd working hours.

Publisher's Acknowledgements

We are grateful to the following for permission to reproduce copyright material:

Figures

Figures 2.1 and 2.12 from Yelp!, http://www.yelp.co.uk/, Yelp! Inc.; Figure 2.3 from NextBio.com; Figures 2.5, 4.13b, 11.10c, 11.11a and 11.13 from Google screenshots, www.google.co.uk; Figure 2.6 from http://biosearch.berkerley.edu, copyright M. A. Hearst; Figure 2.7 from Microsoft product screenshots reprinted with permission from Microsoft Corporation; Figure 2.13 from Findex, copyright (c) 2010 FindEx.com, Inc., and its licensors; Figure 2.15 from Graphical query specification and dynamic result previews for a digital library, *Proceedings of the 11th Annual ACM Symposium on User Interface Software and Technology (UIST'98)* pp. 143–51 (Jones, S. 1998), http://doi.acm.org/10.1145/288392.288595, © 1998 Association for Computing Machinery, Inc. Reprinted by permission; Figure 2.16 from Research: TileBars, http://people.ischool.berkeley.edu/~hearst/research/tilebars.html, copyright M. A. Hearst; Figure 2.17a from *Search User Interfaces*, Cambridge University Press (Hearst, M.A. 2009) figure 10.17a, copyright M. A. Hearst; Figure 2.17b from INSYDER: a content-based visual-information-seeking system for the web, *International Journal on Digital Libraries*, pp. 25–41 (Reiterer, H., Tullius, G. and Mann, T.M. 2005), with kind permission from Springer Science + Business Media and CCC, and Professor H. Reiterer; Figure 2.18 from Using thumbnails to search the Web, *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'01)*, pp. 198–205 (Woodruff, A., Faulring, A., Rosenholtz, R., Morrison, J. and Pirolli, P. 2001), http://doi.acm.org/10.1145/365024.365098, © 2001 Association for Computing Machinery, Inc. Reprinted by permission; Figure 2.20a from Evaluating a system for interactive exploration of large, hierarchically structured document repositories, *Proceedings of the IEEE Symposium on Information Visualization (INFOVIS'04)*, pp. 127–33 (Granitzer, M., Kienreich, W., Sabol, V., Andrews, K. and Klieber, W. 2004), © 2004 IEEE; Figure 2.20b from Search result visualisation with xFIND, *Proceedings of User Interfaces to Data Intensive Systems (UIDIS 2001)*, pp. 50–8 (Andrews, K., Gutl, C., Moser, J., Sabol, V. and Lackner, W. 2001), © 2001 IEEE; Figure 2.21 from http://kylescholz.com/projects/wordnet/, Kyle Scholz;

Figure 2.22 from The Word tree, an interactive visual concordance, *IEEE Transactions on Visualization and Computer Graphics*, 14(6), pp. 1221–8 (Wattenberg, M. and Fernanda, B. 2008), © 2008 IEEE; Figure 2.23 from Baby names popularity graph NameVoyager, http://www.babynamewizard.com; Figure 2.24 from Avian flu case study with nSPace and GeoTime, *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology (VAST'06)* pp. 27–34 (Proulx, P. *et al.* 2006), © 2006 IEEE; Figure 5.4 after Evaluating the accuracy of implicit feedback from clicks and query reformulations in Web search, *ACM Transactions on Information Systems*, 25(2) (Joachims, T., Granka, L., Pan, B., Hembrooke, H., Radlinski, F. and Gay, G. 2007), http://doi.acm.org/10.1145/1229179.1229181, © 2007 Association for Computing Machinery, Inc. Reprinted by permission; Figures 7.4 and 7.5 from The impact of caching on search engines, *Proceedings of the 30th International ACM SIGR Conference on Research and Development in Information Retrieval (SIGI'07)* (Baeza-Yates, R. *et al.* 2007), http://doi.acm.org/10.1145/1277741.1277775, © 2007 Association for Computing Machinery, Inc. Reprinted by permission; Figure 7.6 from Query usage mining in search engines, *Web Mining Applications and Techniques* (Baeza-Yates, R. (Scime, A. ed.) 2004), Idea Group, reprinted by permission of the publisher, IGI Global; Figure 10.1 adapted from Load balancing for term-distributed parallel retrieval, *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 348–355 (Moffat, A., Webber, W. and Zobel, J. 2006), http://doi.acm.org/10.1145/1148170.1148232, © 2006 Association for Computing Machinery, Inc. Reprinted by permission; Figures 10.12 and 10.13 from Challenges on distributed web retrieval, *Proceedings of ICDE 2007*, pp. 6–20 (2007), © 2007 IEEE; Figure 10.14 from A pipelined architecture for distributed text query evaluation, *Information Retrieval*, 10(3), pp. 205–31 (Webber, W., Moffat, A., Zobel, J. and Baeza-Yates, R. 2007), with kind permission from Springer Science + Business Media; Figure 11.1 from Graph structure in the web: experiments and models, *Proceedings of the North Conference on World Wide Web*, pp. 309–20 (Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopolan, S., Stata, R., Tomkins, A. and Wiener, J. 2000), copyright Elsevier (2000); Figure 11.3a from M. Crovella, 1998; Figure 11.3b from Self-similarity in World Wide Web traffic: evidence and possible causes, *SIGMETRICS'96: Proceedings of the 1996 ACM SIGMETRICS International Conference on Measurement and Modelling of Computer Systems*, 24, pp. 160–9 (Crovella, M.E. and Bestavros, A. 1996), http://doi.acm.org/10.1109/90.650143, © 1996 Association for Computing Machinery, Inc. Reprinted by permission; Figures 11.4 and 11.5 from Generic damping functions for propagating importance in link-based ranking algorithms, *Internet Mathematics*, 3(4), pp. 445–78 (Baeza-Yates, R., Boldi, P. and Castillo, C. 2006), copyright 2006 A. K. Peters, Ltd.; Figure 11.7 after Challenges in building large-scale information retrieval systems: invited talk presentation, http://research.google.com/people/jeff/WSDM09-keynote.pdf, Jeffrey Dean; Figure 11.8 from Design trade-offs for search engine caching, *TWEB*, 2(4) (Baeza-Yates, R.A., Gionis, A., Juncqueira, F., Murdock, V., Plachouras, V. and Silvestri, F. 2008), http://doi.acm.org/10.1145/1409220.1409223, © 2008 Association for Computing Machinery, Inc. Reprinted by permission; Figure 11.10a, Ask screenshot, © IAC Search & Media, Inc. 2010. All rights reserved. ASK.COM, ASK JEEVES, the ASK logo, the ASK JEEVES logo and other trade marks appearing on the Ask.com and Ask Jeeves websites are property of IAC Search & Media, Inc. and/or its licensors; Figures 11.10b and 11.15 from Bing screenshots, reprinted with permission from Microsoft

Corporation; Figure 12.8 from Synchronizing a database to improve freshness, *Proceedings of ACM International Conference on Management of Data (SIGMOD)*, pp. 117–28 (Cho, J. and Garcia-Molina, H. 2000), http://doi.acm.org/10.1145/342009.335391, © 2000 Association for Computing Machinery, Inc. Reprinted by permission; Figure 13.9 from INEX 2006 assessment interface, Professor Mounia Lalmas; Figure 14.4 from IBM Almaden Research Center; Figures 14.6 and 14.8 are courtesy of Jim Hafner from the QBIC system, IBM Almaden Research Center; Figure 14.9 from A bipartite graph model for associating images and text, *IJCAI-2007 Workshop on Multimodal Information Retrieval* (Srinivasan, S.H. and Slaney, M. 2007); Figure 14.10 from Image retrieval on large-scale image databases, *Proceedings of the 6th ACM International Conference on Image and Video Retrieval (CIVR 07)*, pp. 17–24 (Horster, E., Lienhart, R. and Slaney, M. 2007), http://doi.acm.org/10.1145/1282280.1282283, © 2007 Association for Computing Machinery, Inc. Reprinted by permission; Figures 14.13 and 14.14 from Kyogu Lee; Figure 14.16 from Video skimming for quick browsing based on audio and image characterization, *Technical Report CMU-CS-95-186* (Smith, M.A. and Kanade, T. 1995), School of Computer Science Tech Report, Carnegie Mellon University; Figure 14.17 from Video manga: generating semantically meaningful video summaries, *MULTIMEDIA '99: Proceedings of the Seventh ACM International Conference on Multimedia (Part 1)*, pp. 383–92 (Uchihashi, S. *et al.* 1999), http://doi.acm.org/10.1145/319463.319654, © 1999 Association for Computing Machinery, Inc. Reprinted by permission; Figure 14.18 from Harpreet Sawhney, Sarnoff Corporation; Figure 14.19 from Salient stills, *ACM Transactions on Multimedia Computing, Communications and Applications*, 1(1), pp. 16–36 (Teodosio, L. and Bender, W. 2005), http://doi.acm.org/10.1145/1047936.1047940, © 2005 Association for Computing Machinery, Inc. Reprinted by permission; Figure 14.20 from PanoramaExcerpts: Extracting and packing panoramas for video browsing, *MULTIMEDIA '97: Proceedings of the Fifth ACM International Conference on Multimedia*, pp. 427–36 (Taniguchi, Y., Akutsu, A. and Tonomura, Y. 1997), http://doi.acm.org/10.1145/266180.266396, © 1997 Association for Computing Machinery, Inc. Reprinted by permission; Figure 14.21 from Hierarchical brushing in a collection of video data, *Proceedings of Hawaii International Conference on System Science (HICSS)* (2001), © 2001, IEEE; Figure 14.26 from Automatic recognition of audio-visual speech: recent progress and challenges, *Proceedings of the IEEE* (Potamianos, G., Neti, C., Gravier, G., Garg, A. and Senior, A.W. 2003), © 2003 IEEE; Figure 14.28 from Multimedia edges: finding hierarchy in all dimensions, *Proceedings of 9th ACM International Conference on Multimedia* (Slaney, M., Ponceleon, D. and Kaufman, J. 2001), http://doi.acm.org/10.1145/500141.500149, © 2001 Association for Computing Machinery, Inc. Reprinted by permission; Figure 14.29 from Comparison of automatic shot boundary detection algorithms, *SPIE Image and Video Processing VII*, pp. 3656–729 (Lienhart, R. 1999), SPIE; Figure 15.3 from Oxfam Australia; Figure 15.5 from Evaluation by comparing result sets in context, *Proceedings of the 15th ACM International Conference on Information and Knowledge Management* pp. 94–101 (Thomas, P. and Hawking, D. 2006), http://doi.acm.org/10.1145/1183614.1183632, © ACM, 2006; Figure 16.1 from Edie Rasmussen, and with permission from The Network Development and MARC Standards Office; Figure 16.2 from Find... books or journals, http://www.library.ubc.ca/home/research.html, The University of British Columbia Library website (2010). Used with permission; Figures 16.4, 16.5, 16.6 and 16.7 from DIALOG, Dialog® interface and screen shots reproduced with permission of Dialog,

LLC. The Dialog product name is a registered trademark of Dialog, LLC; Figure 16.4 with permission from EBSCO Publishing, Inc.

Tables

Table 4.2 after Overview of the sixth text retrieval conference (TREC- 6), *Proceedings of the Sixth Text REtrieval Conference (TREC-6)* (Voorhees, E. and Harman, D. 1997); Table 7.3 from From e-sex to e-commerce: Web search changes, *Computer*, 35(3), pp. 107–9 (Spink, A., Jansen, B.J., Wolfram, D. and Saracevic, T. 2002), © 2002 IEEE.

Text

Extract on page 159 from http://trec.nist.gov, NIST.

In some instances we have been unable to trace the owners of copyright material, and we would appreciate any information that would enable us to do so.