Modern Information Retrieval

the concepts and technology behind search

Second edition

Ricardo Baeza-Yates
Berthier Ribeiro-Neto

Modern Information Retrieval

The Concepts and Technology behind Search

Ricardo Baeza-Yates
Berthier Ribeiro-Neto

Second edition

Addison-Wesley

Harlow, England • Reading, Massachusetts
Menlo Park, California • New York
Don Mills, Ontario • Amsterdam • Bonn
Sydney • Singapore • Tokyo • Madrid
San Juan • Milan • Mexico City • Seoul • Taipei

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 four 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.
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, and personalization, to mention just some of them.
5. An improved Web site 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

Author's Acknowledgements to the Second Edition

We would like to sincerely thank for 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 remains, hopefully only few, are our entirely 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?!

Author's Acknowledgements to the First Edition

We would like to sincerely thank for 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.