Christopher D. Manning
Prabhakar Raghavan
Hinrich Schütze

Introduction to Information Retrieval

CAMBRIDGE

CAMBRIDGE
UNIVERSITY PRESS

4381/4 Ansari Road, Daryaganj, Delhi 110002, India

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of education, learning and research at the highest international levels of excellence.

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

First published 2008
First South Asian edition 2012
Reprinted 2013, 2014

This South Asian edition is based on Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze / Introduction to Information Retrieval / 9780521865715 / 2008

Printed in India by Shree Maitrey Printech Pvt. Ltd., Noida

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-13 978-1-107-66639-9 (paperback)

Contents

Zhang, Jiangong, Xiaohui Long, and Torsten Suel. 2007. Performance of compressed inverted list caching in search engines. In *Proc. CIKM*. ACM Press. [98]

Zhang, Tong, and Frank J. Oles. 2001. Text categorization based on regularized linear classification methods. *IR* 4(1):5–31. URL: citeseer.ist.psu.edu/zhang00text.html. [319]

Zhao, Ying, and George Karypis. 2002. Evaluation of hierarchical clustering algorithms for document datasets. In *Proc. CIKM*, pp. 515–524. ACM Press. DOI: http://doi.acm.org/10.1145/584792.584877. [367]

Zipf, George Kingsley. 1949. *Human Behavior and the Principle of Least Effort*. Addison-Wesley. [97]

Zobel, Justin. 1998. How reliable are the results of large-scale information retrieval experiments? In *Proc. SIGIR*, pp. 307–314. [160]

Zobel, Justin, and Philip Dart. 1995. Finding approximate matches in large lexicons. *Software Practice and Experience* 25(3):331–345. URL: citeseer.ifi.unizh.ch/zobel95finding.html. [60]

Zobel, Justin, and Philip Dart. 1996. Phonetic string matching: Lessons from information retrieval. In *Proc. SIGIR*, pp. 166–173. ACM Press. [60]

Zobel, Justin, and Alistair Moffat. 2006. Inverted files for text search engines. *ACM Computing Surveys* 38(2). [17, 76, 98, 122]

Zobel, Justin, Alistair Moffat, Ross Wilkinson, and Ron Sacks-Davis. 1995. Efficient retrieval of partial documents. *IP&M* 31(3):361–377. DOI: http://dx.doi.org/10.1016/0306-4573(94)00052-5. [199]

Zukowski, Marcin, Sandor Heman, Niels Nes, and Peter Boncz. 2006. Super-scalar RAM-CPU cache compression. In *Proc. International Conference on Data Engineering*, p. 59. IEEE Computer Society. DOI: http://dx.doi.org/10.1109/ICDE.2006.150. [98]

Index

Turpin, Andrew, and William R. Hersh. 2001. Why batch and user evaluations do not give the same results. In *Proc. SIGIR*, pp. 225–231.

Turpin, Andrew, and William R. Hersh. 2002. User interface effects in past batch versus user experiments. In *Proc. SIGIR*, pp. 431–432.

Turpin, Andrew, Yohannes Tsegay, David Hawking, and Hugh E. Williams. 2007. Fast generation of result snippets in web search. In *Proc. SIGIR*, pp. 127–134. ACM Press. [161]

Turtle, Howard. 1994. Natural language vs. Boolean query evaluation: A comparison of retrieval performance. In *Proc. SIGIR*, pp. 212–220. ACM Press. [15]

Turtle, Howard, and W. Bruce Croft. 1989. Inference networks for document retrieval. In *Proc. SIGIR*, pp. 1–24. ACM Press. [215]

Turtle, Howard, and W. Bruce Croft. 1991. Evaluation of an inference network-based retrieval model. *TOIS* 9(3):187–222. [215]

Turtle, Howard, and James Flood. 1995. Query evaluation: strategies and optimizations. *IP&M* 31(6):831–850. DOI: http://dx.doi.org/10.1016/0306-4573(95)00020-H. [123]

Vaithyanathan, Shivakumar, and Byron Dom. 2000. Model-based hierarchical clustering. In *Proc. UAI*, pp. 599–608. Morgan Kaufmann. [368]

van Rijsbergen, C. J. 1979. *Information Retrieval*, 2nd edition. Butterworths. [159, 198, 203, 213, 216]

van Rijsbergen, C. J. 1989. Towards an information logic. In *SIGIR*, pp. 77–86. ACM Press. DOI: http://doi.acm.org/10.1145/75334.75344. [xviii]

van Zwol, Roelof, Jeroen Baas, Herre van Oostendorp, and Frans Wiering. 2006. Bricks: The building blocks to tackle query formulation in structured document retrieval. In *Proc. ECIR*, pp. 314–325. [200]

Vapnik, Vladimir N. 1998. *Statistical Learning Theory*. Wiley-Interscience. [319]

Vittaut, Jean-Noël, and Patrick Gallinari. 2006. Machine learning ranking for structured information retrieval. In *Proc. ECIR*, pp. 338–349. [199]

Voorhees, Ellen M. 1985a. The cluster hypothesis revisited. In *Proc. SIGIR*, pp. 188–196. ACM Press. [344]

Voorhees, Ellen M. 1985b. *The effectiveness and efficiency of agglomerative hierarchic clustering in document retrieval*. Technical Report TR 85-705, Cornell. [367]

Voorhees, Ellen M. 2000. Variations in relevance judgments and the measurement of retrieval effectiveness. *IP&M* 36:697–716. [160]

Voorhees, Ellen M., and Donna Harman (eds.). 2005. *TREC: Experiment and Evaluation in Information Retrieval*. MIT Press. [159, 453, 461]

Wagner, Robert A., and Michael J. Fischer. 1974. The string-to-string correction problem. *JACM* 21(1):168–173. DOI: http://doi.acm.org/10.1145/321796.321811. [59]

Ward Jr., J. H. 1963. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association* 58:236–244. [367]

Wei, Xing, and W. Bruce Croft. 2006. LDA-based document models for ad-hoc retrieval. In *Proc. SIGIR*, pp. 178–185. ACM Press. DOI: http://doi.acm.org/10.1145/1148170.1148204. [384]

Weigend, Andreas S., Erik D. Wiener, and Jan O. Pedersen. 1999. Exploiting hierarchy in text categorization. *IR* 1(3):193–216. [319]

Weston, Jason, and Chris Watkins. 1999. Support vector machines for multi-class pattern recognition. In *Proc. European Symposium on Artificial Neural Networks*, pp. 219–224. [319]

Williams, Hugh E., and Justin Zobel. 2005. Searchable words on the web. *International Journal on Digital Libraries* 5(2):99–105. DOI: http://dx.doi.org/10.1007/s00799-003-0050-z. [97]

Williams, Hugh E., Justin Zobel, and Dirk Bahle. 2004. Fast phrase querying with combined indexes. *TOIS* 22(4):573–594. [41]

Witten, Ian H., and Timothy C. Bell. 1990. Source models for natural language text. *International Journal Man-Machine Studies* 32(5):545–579. [97]

Witten, Ian H., and Eibe Frank. 2005. *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd edition. Morgan Kaufmann. [342]

Witten, Ian H., Alistair Moffat, and Timothy C. Bell. 1999. *Managing Gigabytes: Compressing and Indexing Documents and Images*, 2nd edition. Morgan Kaufmann. [76, 97, 98]

Wong, S. K. Michael, Yiyu Yao, and Peter Bollmann. 1988. Linear structure in information retrieval. In *Proc. SIGIR*, pp. 219–232. ACM Press. [320]

Woodley, Alan, and Shlomo Geva. 2006. NLPX at INEX 2006. In *Proc. INEX*, pp. 302–311. [200]

Xu, Jinxi, and W. Bruce Croft. 1996. Query expansion using local and global document analysis. In *Proc. SIGIR*, pp. 4–11. ACM Press. [177]

Xu, Jinxi, and W. Bruce Croft. 1999. Cluster-based language models for distributed retrieval. In *Proc. SIGIR*, pp. 254–261. ACM Press. DOI: http://doi.acm.org/10.1145/312624.312687. [344]

Yang, Hui, and Jamie Callan. 2006. Near-duplicate detection by instance-level constrained clustering. In *Proc. SIGIR*, pp. 421–428. ACM Press. DOI: http://doi.acm.org/10.1145/1148170.1148243. [344]

Yang, Yiming. 1994. Expert network: Effective and efficient learning from human decisions in text categorization and retrieval. In *Proc. SIGIR*, pp. 13–22. ACM Press. [291]

Yang, Yiming. 1999. An evaluation of statistical approaches to text categorization. *IR* 1:69–90. [319]

Yang, Yiming. 2001. A study of thresholding strategies for text categorization. In *Proc. SIGIR*, pp. 137–145. ACM Press. DOI: http://doi.acm.org/10.1145/383952.383975. [292]

Yang, Yiming, and Bryan Kisiel. 2003. Margin-based local regression for adaptive filtering. In *Proc. CIKM*, pp. 191–198. ACM Press. DOI: http://doi.acm.org/10.1145/956863.956902. [292]

Yang, Yiming, and Xin Liu. 1999. A re-examination of text categorization methods. In *Proc. SIGIR*, pp. 42–49. ACM Press. [265, 319]

Yang, Yiming, and Jan Pedersen. 1997. Feature selection in statistical learning of text categorization. In *Proc. ICML*. [265]

Yue, Yisong, Thomas Finley, Filip Radlinski, and Thorsten Joachims. 2007. A support vector method for optimizing average precision. In *Proc. SIGIR*. ACM Press. [320]

Zamir, Oren, and Oren Etzioni. 1999. Grouper: A dynamic clustering interface to web search results. In *Proc. WWW*, pp. 1361–1374. Elsevier North-Holland. DOI: http://dx.doi.org/10.1016/S1389-1286(99)00054-7. [344, 368]

Zaragoza, Hugo, Djoerd Hiemstra, Michael Tipping, and Stephen Robertson. 2003. Bayesian extension to the language model for ad hoc information retrieval. In *Proc. SIGIR*, pp. 4–9. ACM Press. [232]

Zavrel, Jakub, Peter Berck, and Willem Lavrijssen. 2000. Information extraction by text classification: Corpus mining for features. In *Proc. Workshop Information Extraction meets Corpus Linguistics*. URL: http://www.cnts.ua.ac.be/Publications/2000/ZBL00. Held in conjunction with LREC-2000. [292]

Zha, Hongyuan, Xiaofeng He, Chris H. Q. Ding, Ming Gu, and Horst D. Simon. 2001. Bipartite graph partitioning and data clustering. In *Proc. CIKM*, pp. 25–32. ACM Press. [345, 368]

Zhai, Chengxiang, and John Lafferty. 2001a. Model-based feedback in the language modeling approach to information retrieval. In *Proc. CIKM*. ACM Press. [231]

Zhai, Chengxiang, and John Lafferty. 2001b. A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proc. SIGIR*, pp. 334–342. ACM Press. [232]

Zhai, ChengXiang, and John Lafferty. 2002. Two-stage language models for information retrieval. In *Proc. SIGIR*, pp. 49–56. ACM Press. DOI: http://doi.acm.org/10.1145/564376.564387. [233]

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.

Cover design by David Levy

CAMBRIDGE
UNIVERSITY PRESS
www.cambridge.org

ISBN 978-1-107-66639-9