Keith van Rijsbergen

The Geometry of Information Retrieval

CAMBRIDGE

The Geometry of Information Retrieval

Information retrieval, IR, is the science of extracting information from documents. It can be viewed in a number of ways: logical, probabilistic and vector space models are some of the most important. In this book, the author, one of the leading researchers in the area, shows how these three views can be combined in one mathematical framework, the very one used to formulate the general principles of quantum mechanics. Using this framework, van Rijsbergen presents a new theory for the foundations of IR, in particular a new theory of measurement. He shows how a document can be represented as a vector in Hilbert space, and the document's relevance by an Hermitian operator. All the usual quantum-mechanical notions, such as uncertainty, superposition and observable, have their IR-theoretic analogues. But the approach is more than just analogy: the standard theorems can be applied to address problems in IR, such as pseudo-relevance feedback, relevance feedback and ostensive retrieval. The relation with quantum computing is also examined. To help keep the book self-contained, appendices with background material on physics and mathematics are included, and each chapter ends with some suggestions for further reading. This is an important book for all those working in IR, AI and natural language processing.

KEITH VAN RIJSBERGEN'S research has, since 1969, been devoted to information retrieval, working on both theoretical and experimental aspects. His current research is concerned with the design of appropriate logics to model the flow of information and the application of Hilbert space theory to content-based IR. This is his third book on IR: his first is now regarded as the classic text in the area. In addition he has published over 100 research papers and is a regular speaker at major IR conferences. Keith is a Fellow of the IEE, BCS, ACM, and the Royal Society of Edinburgh. In 1993 he was appointed Editor-in-Chief of *The Computer Journal*, an appointment he held until 2000. He is an associate editor of *Information Processing and Management*, on the editorial board of *Information Retrieval*, and on the advisory board of the *Journal of Web Semantics*. He has served as a programme committee member and editorial board member of the major IR conferences and journals. He is a non-executive director of a start-up: Virtual Mirrors Ltd.

The Geometry of Information Retrieval

C. J. VAN RIJSBERGEN

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
The Pitt Building, Trumpington Street, Cambridge, United Kingdom

Cambridge University Press
The Edinburgh Building, Cambridge, CB2 8RU, UK

Published in the United States of America by Cambridge University Press, New York

www.cambridge.org
Information on this title: www.cambridge.org/9780521838054

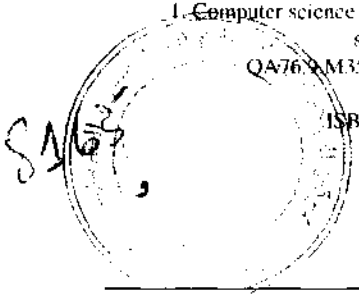
First published 2004
Fourth printing 2007

Printed in the United Kingdom at the University Press, Cambridge

A catalogue record for this publication is available from the British Library

Library of Congress Cataloguing in Publication data
Van Rijsbergen, C. J., 1943–
The geometry of information retrieval / by C. J. van Rijsbergen.
p. cm.
Includes bibliographical references and index.
ISBN 0 521 83805 3 (hb)
1. Computer science – Mathematics. 2. Information storage and retrieval
systems – Mathematics. I. Title.
QA76.9.M35.V38 2004 025.04 dc22 2004045683

ISBN-13 978-0-521-83805-4 hardback



To make a start,
Out of particulars
And make them general, rolling
Up the sum, by defective means

Paterson: Book I
William Carlos Williams, 1992

for
Nicola

Contents