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

PUBLISHED BY THE PRESS SYNDICATE OF THE UNIVERSITY OF CAMBRIDGE
The Pitt Building, Trumpington Street, Cambridge, United Kingdom

CAMBRIDGE UNIVERSITY PRESS
The Edinburgh Building, Cambridge, CB2 2RU, UK
40 West 20th Street, New York, NY 10011–4211, USA
477 Williamstown Road, Port Melbourne, VIC 3207, Australia
Ruiz de Alarcón 13, 28014 Madrid, Spain
Dock House, The Waterfront, Cape Town 8001, South Africa

http://www.cambridge.org

First published 2004

Printed in the United Kingdom at the University Press, Cambridge

Typeface Times 10/13 pt. *System* LaTeX 2ε [TB]

A catalogue record for this book 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 0 521 83805 3 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

Preface

This book begins and ends in information retrieval, but travels through a route constructed in an abstract way. In particular it goes through some of the most interesting and important models for information retrieval, a vector space model, a probabilistic model and a logical model, and shows how these three and possibly others can be described and represented in Hilbert space. The reasoning that occurs within each one of these models is formulated algebraically and can be shown to depend essentially on the geometry of the information space. The geometry can be seen as a 'language' for expressing the different models of information retrieval.

The approach taken is to structure these developments firmly in terms of the mathematics of Hilbert spaces and linear operators. This is of course the approach used in quantum mechanics. It is remarkable that the application of Hilbert space mathematics to information retrieval is very similar to its application to quantum mechanics. A document in IR can be represented as a vector in Hilbert space, and an observable such as 'relevance' or 'aboutness' can be represented by a Hermitian operator. However, this is emphatically not a book about quantum mechanics but about using the same language, the mathematical language of quantum mechanics, for the description of information retrieval. It turns out to be very convenient that quantum mechanics provides a ready-made interpretation of this language. It is as if in physics we have an example semantics for the language, and as such it will be used extensively to motivate a similar but different interpretation for IR. We introduce an appropriate logic and probability theory for information spaces guided by their introduction into quantum mechanics. Gleason's Theorem, which specifies an algorithm for computing probabilities associated with subspaces in Hilbert space, is of critical importance in quantum mechanics and will turn out to be central for the same reasons in information retrieval. Whereas quantum theory is about a theory of measurement for *natural* systems, The Geometry of Information Retrieval is about

such a theory for *artificial* systems, and in particular for information retrieval. The important notions in quantum mechanics, state vector, observable, uncertainty, complementarity, superposition and compatibility readily translate into analogous notions in information retrieval, and hence the theorems of quantum theory become available as theorems in IR.

One of the main aims of this book is to present the requisite mathematics to explore in detail the foundation of information retrieval as a parallel to that of quantum mechanics. The material is principally addressed to students and researchers in information retrieval but will also be of interest to those working in such disciplines as AI and quantum computation. An attempt is made to lay a sound mathematical foundation for reasoning about existing models in IR sufficient for their modification and extension. The hope is that the treatment will inspire and enable the invention of new models. All the mathematics is introduced in an elementary fashion, step-by-step, making copious references to matching developments in quantum mechanics. Any reader with a good grasp of high school mathematics, or A-level equivalent, should be able to follow the mathematics from first principles. One exception to this is the material in the Prologue, where some more advanced notions are rapidly introduced, as is often the case in dialogue, but even there a quick consultation of the appropriate appendices would clarify the mathematics.

Although the material is not about quantum computation, it could easily be adopted as an elementary introduction to that subject. The mathematics required to understand most discussions on quantum computation is covered. It will be interesting to see if the approach taken to modelling IR can be mapped onto a quantum computer architecture. In the quantum computation literature the Dirac notation is used as a *lingua franca*, and it is also used here and is explained in some detail as it is needed.

Students and researchers in IR are happy to use mathematics to define and specify algorithms to implement sophisticated search strategies, but they seem to be notoriously resistant to investing energy and effort into acquiring new mathematics. Thus there is a threshold to be overcome in convincing a person to take the time to understand the mathematics that is here. For this reason we begin with a Prologue. In it fundamental concepts are presented and discussed with only a little use of mathematics, to introduce by way of a dialogue the new way of thinking about IR. It is hoped that illustrating the material in this way will overcome some of the reader's resistance to venturing into this new mathematical territory for IR.

A further five chapters followed by three technical appendices and an extensive annotated Bibliography constitute the full extent of the book. The chapters make up a progression. Chapter 1, the Introduction, goes some way to showing

the extent to which the material depends on ideas from quantum mechanics whilst at the same time motivating the shift in thinking about IR notions. Chapter 2 gives an account of traditional Boolean algebra based on set theory and shows how non-Boolean structures arise naturally when classes are no longer sets, but are redefined in an appropriate way. An illustration of the breakdown of the law of distribution in logic then gives rise to non-classical logic. Chapter 3 introduces vector and Hilbert spaces from first principles, leading to Chapter 4 which describes linear operators, their representation and properties as vehicles for measurement and observation. Chapter 5 is the first serious IR application for the foregoing theory. It builds on the earlier work of many researchers on logics for IR and it shows how conditionals in logic can be represented as objects in Hilbert space. Chapter 6, by far the longest, takes the elementary theory presented thus far and recasts it, using the Dirac notation, so that it can be applied to a number of specific problems in IR, for example, pseudo-relevance feedback, relevance feedback and ostensive retrieval.

Each chapter concludes with some suggestions for further reading, thus providing guidance for possible extensions. In general the references collected at the end of the book are extensively annotated. One reason for this is that readers, not necessarily acquainted with quantum mechanics or its mathematics, may enjoy further clarification as to why pursuing any further reference may be worthwhile. Scanning the bibliography with its annotations is intended to provide useful information about the context for the ideas in the book. A given reference may refer to a number of others because they relate to the same topic, or provide a commentary on the given one.

There are three detailed appendices. The first one gives a potted introduction to linear algebra for those who wish to refresh their memories on that subject. It also conveniently contains a summary of the Dirac notation which takes some getting used to. The second appendix is a self-contained introduction to quantum mechanics, and it uses the Dirac notation explained in the previous appendix. It also contains a simple proof of the Heisenberg Uncertainty Principle which does not depend on any physics. The final appendix gives the classical axioms for probability theory and shows how they are extended to quantum probability.

There a number of ways of reading this book. The obvious way is to read it from beginning to end, and in fact it has been designed for that. Another way is to read the Prologue, the Introduction and the appendices, skipping the intervening chapters on a first pass; this would give the reader a conceptual grasp of the material without a detailed understanding of the mathematics. A third way is to read the Prologue last, and then the bulk of the book will provide grounding for some of the advanced mathematical ideas that are introduced

rapidly in the Prologue. One can also skip all the descriptive and motivational material and start immediately with the mathematics, for that one begins at Chapter 2, and continues to the end. A fifth way is to read only Chapter 6, the geometry of IR, and consult the relevant earlier chapters as needed.

There are many people who have made the writing of this book possible. Above all I would like to thank Juliet and Nicola van Rijsbergen for detailed and constructive comments on earlier drafts of the manuscripts, and for the good humour with which they coped with my frustrations. Mounia Lalmas, Thomas Roelleke and Peter Bruza I thank for technical comments on an early draft. Elliott Sober I thank for help with establishing the origin of some of the quotations as well as helping me clarify some thinking. Dealing with a publisher can sometimes be fraught with difficulties; fortunately David Tranah of CUP ensured that it was in fact wonderfully straightforward and agreeable, for which I express my appreciation; I also thank him for his constant encouragement. The ideas for the monograph were conceived during 2000–1 whilst I was on sabbatical at Cambridge University visiting the Computer Laboratory, Department of Engineering and King's College, all of which institutions deserve thanks for hosting me and making it possible to think and write. Taking on a task such as this inevitably means that less time is available for other things, and here I would like to express my appreciation to the IR group at Glasgow University for their patience. Finally, I would like record my intellectual debt to Bill Maron whose ideas in many ways foreshadowed some of mine, also to the writings of John von Neumann for his insights on geometry, logic and probability without which I could not have begun.