Information Retrieval

Second Edition

C. J. van Rijsbergen

Butterworths

INFORMATION RETRIEVAL

Second Edition

C. J. van Rijsbergen B.Sc. Ph.D. M.B.C.S.

Computer Laboratory,
University of Cambridge

BUTTERWORTHS

London Boston

Sydney Durban Wellington Toronto

United Kingdom	**Butterworth & Co (Publishers) Ltd**
London	88 Kingsway, WC2B 6AB
Australia	**Butterworths Pty Ltd**
Sydney	586 Pacific Highway, Chatswood, NSW 2067
	Also at Melbourne, Brisbane, Adelaide and Perth
Canada	**Butterworth & Co (Canada) Ltd**
Toronto	2265 Midland Avenue, Scarborough, Ontario, M1P 4S1
New Zealand	**Butterworths of New Zealand Ltd**
Wellington	77-85 Customhouse Quay, 1
	T & W Young Building, CPO Box 472
South Africa	**Butterworth & Co (South Africa) Ltd**
Durban	152-154 Gale Street
USA	**Butterworth (Publishers) Inc**
Boston	19 Cummings Park, Woburn, Massachusetts 01801

First published 1975
Second edition 1979

ISBN 0 408 70929 4

British Library Cataloguing in Publication Data

Van Rijsbergen, C. J.
Information retrieval. – 2nd ed.
1. Information storage and retrieval systems
I. Title
029.7 Z699 78-40725

ISBN 0-408-70929-4

Printed in Great Britain by
The Whitefriars Press Ltd, London and Tonbridge

INFORMATION RETRIEVAL

PREFACE TO THE SECOND EDITION

The major change in the second edition of this book is the addition of a new chapter on probabilistic retrieval. This chapter has been included because I think this is one of the most interesting and active areas of research in information retrieval. There are still many problems to be solved so I hope that this particular chapter will be of some help to those who want to advance the state of knowledge in this area. All the other chapters have been updated by including some of the more recent work on the topics covered. In preparing this new edition I have benefited from discussions with Bruce Croft, David Harper, Stephen Robertson and Karen Sparck Jones. I am grateful to the University of Cambridge Computer Laboratory for providing me with the facilities for carrying out the work. Finally, I am indebted to the Royal Society for supporting me on their Scientific Information Research Fellowship.

C.J.v.R.

PREFACE TO THE FIRST EDITION

The material of this book is aimed at advanced undergraduate information (or computer) science students, postgraduate library science students, and research workers in the field of IR. Some of the chapters, particularly Chapter 6*, make *simple* use of a little advanced mathematics. However, the necessary mathematical tools can be easily mastered from numerous mathematical texts that now exist and in any case references have been given where the mathematics occur.

I had to face the problem of balancing clarity of exposition with density of references. I was tempted to give large numbers of references but was afraid they would have destroyed the continuity of the text. I have tried to steer a middle course and not compete with the *Annual Review of Information Science and Technology*.

Normally one is encouraged to cite only works that have been published in some readily accessible form such as a book or periodical. Unfortunately much of the interesting work in IR is contained in technical reports and Ph.D. theses. For example most of the work done on the SMART system at Cornell is available only in reports. Luckily many of these are now available through the National Technical Information Service (U.S.) and University Microfilms (U.K.). I have not avoided using these sources although if the same material is accessible more readily in some other form I have given it preference.

I should like to acknowledge my considerable debt to many people and institutions that have helped me. Let me say first that they are responsible for many of the ideas in this book but that only I wish to be held responsible. My greatest debt is to Karen Sparck Jones who

* This is Chapter 7 in the second edition.

taught me to research information retrieval as an experimental science. Nick Jardine and Robin Sibson taught me about the theory of automatic classification. Cyril Cleverdon is responsible for forcing me to think about evaluation. Mike Keen helped by providing data. Gerry Salton has influenced my thinking about IR considerably, mainly through his published work. Ken Moody had the knack of bailing me out when the going was rough and encouraging me to continue experimenting. Juliet Gundry is responsible for making the text more readable and clear. Bruce Croft, who read the final draft, made many useful comments. Ness Barry takes all the credit for preparing the manuscript. Finally, I am grateful to the Office of Scientific and Technical Information for funding most of the early experimental work on which the book is based; to the King's College Research Centre for providing me with an environment in which I could think, and to the Department of Information Science at Monash University for providing me with the facilities for writing.

C.J.v.R

CONTENTS

One
INTRODUCTION

Information retrieval is a wide, often loosely-defined term but in these pages I shall be concerned only with automatic information retrieval systems. Automatic as opposed to manual and information as opposed to data or fact. Unfortunately the word information can be very misleading. In the context of information retrieval (IR), information, in the technical meaning given in Shannon's theory of communication, is not readily measured (Shannon and Weaver[1]). In fact in many cases, one can adequately describe the kind of retrieval by simply substituting 'document' for 'information'. Nevertheless, 'information retrieval' has become accepted as a description of the kind of work published by Cleverdon, Salton, Sparck Jones, Lancaster and others. A perfectly straightforward definition along these lines is given by Lancaster[2]: 'Information retrieval is the term conventionally, though somewhat inaccurately, applied to the type of activity discussed in this volume. An information retrieval system does not inform (i.e. change the knowledge of) the user on the subject of his inquiry. It merely informs on the existence (or non-existence) and whereabouts of documents relating to his request.' This specifically excludes Question-Answering systems as typified by Winograd[3] and those described by Minsky[4]. It also excludes data retrieval systems such as used by, say, the stock exchange for on-line quotations.

To make clear the difference between data retrieval (DR) and information retrieval (IR), I have listed in Table 1.1 some of the distinguishing properties of data and information retrieval. One may want to criticise this dichotomy on the grounds that the boundary

TABLE 1.1. DATA RETRIEVAL OR INFORMATION RETRIEVAL?

	Data Retrieval (DR)	*Information Retrieval* (IR)
Matching	Exact match	Partial match, best match
Inference	Deduction	Induction
Model	Deterministic	Probabilistic
Classification	Monothetic	Polythetic
Query language	Artificial	Natural
Query specification	Complete	Incomplete
Items wanted	Matching	Relevant
Error response	Sensitive	Insensitive

between the two is a vague one. And so it is, but it is a useful one in that it illustrates the range of complexity associated with each mode of retrieval.

Let us now take each item in the table in turn and look at it more closely. In data retrieval we are normally looking for an exact match, that is, we are checking to see whether an item is or is not present in the file. In information retrieval this may sometimes be of interest but more generally we want to find those items which partially match the request and then select from those a few of the best matching ones.

The inference used in data retrieval is of the simple deductive kind, that is, aRb and bRc then aRc. In information retrieval it is far more common to use inductive inference; relations are only specified with a degree of certainty or uncertainty and hence our confidence in the inference is variable. This distinction leads one to describe data retrieval as deterministic but information retrieval as probabilistic. Frequently Bayes' Theorem is invoked to carry out inferences in IR, but in DR probabilities do not enter into the processing.

Another distinction can be made in terms of the classifications that are likely to be useful. In DR we are most likely to be interested in a monothetic classification, that is, one with classes defined by objects possessing attributes both necessary and sufficient to belong to a class. In IR such a classification is on the whole not very useful, in fact more often a polythetic classification is what is wanted. In such a classification each individual in a class will possess only a proportion of all the attributes possessed by all the members of that class. Hence no attribute is necessary nor sufficient for membership to a class.

The query language for DR will generally be of the artificial kind, one with restricted syntax and vocabulary, in IR we prefer to use natural language although there are some notable exceptions. In DR the query is generally a complete specification of what is wanted, in IR it is invariably incomplete. This last difference arises partly from the fact

that in IR we are searching for relevant documents as opposed to exactly matching items. The extent of the match in IR is assumed to indicate the likelihood of the relevance of that item. One simple consequence of this difference is that DR is more sensitive to error in the sense that, an error in matching will not retrieve the wanted item which implies a total failure of the system. In IR small errors in matching generally do not affect performance of the system significantly,

Many automatic information retrieval systems are *experimental.* I only make occasional reference to *operational* systems. Experimental IR is mainly carried on in a 'laboratory' situation whereas operational systems are commercial systems which charge for the service they provide. Naturally the two systems are evaluated differently. The 'real world' IR systems are evaluated in terms of 'user satisfaction' and the price the user is willing to pay for its service. Experimental IR systems are evaluated by comparing the retrieval experiments with standards specially constructed for the purpose. I believe that a book on *experimental* information retrieval, covering the design and evaluation of retrieval systems from a point of view which is independent of any particular system, will be a great help to other workers in the field and indeed is long overdue.

Many of the tecniques I shall discuss will not have proved themselves incontrovertibly superior to all other techniques, but they have promise and their promise will only be realised when they are understood. Information about new techniques has been so scattered through the literature that to find out about them you need to be an expert before you begin to look. I hope that I will be able to take the reader to the point where he will have little trouble in implementing some of the new techniques. Also, that some people will then go on to experiment with them, and generate new, convincing evidence of their efficiency and effectiveness.

My aim throughout has been to give a complete coverage of the more important ideas current in various special areas of information retrieval. Inevitably some ideas have been elaborated at the expense of others. In particular, emphasis is placed on the use of automatic classification techniques and rigorous methods of measurement of effectiveness. On the other hand, automatic content analysis is given only a superficial coverage. The reasons are straightforward, firstly the material reflects my own bias, and secondly, no adequate coverage of the first two topics has been given before whereas automatic content analysis has been documented very well elsewhere. A subsidiary reason for emphasising automatic classification is that little appears to be known or understood about it in the context of IR so that research workers are loath to experiment with it.

The structure of the book

The introduction presents some basic background material, demarcates the subject and discusses loosely some of the problems in IR. The chapters that follow cover topics in the order in which I would think about them were I about to design an experimental IR system. They begin by describing the generation of machine representations for the information, and then move on to an explanation of the logical structures that may be arrived at by clustering. There are numerous methods for representing these structures in the computer, or in other words, there is a choice of file structures to represent the logical structure, so these are outlined next. Once the information has been stored in this way we are able to search it, hence a discussion of search strategies follows. The chapter on probabilistic retrieval is an attempt to create a formal model for certain kinds of search strategies. Lastly, in an experimental situation all of the above will have been futile unless the results of retrieval can be *evaluated.* Therefore a large chapter is devoted to ways of measuring the effectiveness of retrieval. In the final chapter I have indulged in a little speculation about the possibilities for IR in the next decade.

The two major chapters are those dealing with automatic classification and evaluation. I have tried to write them in such a way that each can be read independently of the rest of the book (although I do not recommend this for the non-specialist).

Outline

Chapter 2: Automatic Text Analysis–contains a straightforward discussion of how the text of a document is represented inside a computer. This is a superficial chapter but I think it is adequate in the context of this book.

Chapter 3: Automatic Classification–looks at automatic classification methods in general and then takes a deeper look at the use of these methods in information retrieval.

Chapter 4: File Structures–here we try and discuss file structures from the point of view of someone primarily interested in information retrieval.

Chapter 5: Search Strategies–gives an account of some search strategies when applied to document collections structured in different ways. It also discusses the use of feedback.

Chapter 6: Probabilistic Retrieval–describes a formal model for enhancing retrieval effectiveness by using sample information about the

frequency of occurrence and co-occurrence of index terms in the relevant and non-relevant documents.
Chapter 7: Evaluation—here I give a traditional view of the measurement of effectiveness followed by an explanation of some of the more promising attempts at improving the art. I also attempt to provide foundations for a theory of evaluation.
Chapter 8: The Future—contains some speculation about the future of IR and tries to pinpoint some areas of research where further work is desperately needed.

Information retrieval

Since the 1940s the problem of information storage and retrieval has attracted increasing attention. It is simply stated: we have vast amounts of information to which accurate and speedy access is becoming ever more difficult. One effect of this is that relevant information gets ignored since it is never uncovered, which in turn leads to much duplication of work and effort. With the advent of computers, a great deal of thought has been given to using them to provide rapid and intelligent retrieval systems. In libraries, many of which certainly have an information storage and retrieval problem, some of the more mundane tasks, such as cataloguing and general administration, have successfully been taken over by computers. However, the problem of effective retrieval remains largely unsolved.

In principle, information storage and retrieval is simple. Suppose there is a store of documents and a person (user of the store) formulates a question (request or query) to which the answer is a set of documents satisfying the information need expressed by his question. He can obtain the set by reading all the documents in the store, retaining the relevant documents and discarding all the others. In a sense, this constitutes 'perfect' retrieval. This solution is obviously impracticable. A user either does not have the time or does not wish to spend the time reading the entire document collection, apart from the fact that it may be physically impossible for him to do so.

When high speed computers became available for non-numerical work, many thought that a computer would be able to 'read' an entire document collection to extract the relevant documents. It soon became apparent that using the natural language text of a document not only caused input and storage problems (it still does) but also left unsolved the intellectual problem of characterising the document content. It is conceivable that future hardware developments may make natural

language input and storage more feasible. But automatic characterisation in which the software attempts to duplicate the human process of 'reading' is a very sticky problem indeed. More specifically 'reading' involves attempting to extract information, both syntactic and semantic, from the text and using it to decide whether each document is relevant or not to a particular request. The difficulty is not only knowing how to extract the information but also how to use it to decide relevance. The comparatively slow progress of modern linguistics on the semantic front and the conspicuous failure of machine translation (Bar-Hillel[5]) show that these problems are largely unsolved.

The reader will have noticed that already, the idea of 'relevance' has slipped into the discussion. It is this notion which is at the centre of information retrieval. The purpose of an automatic retrieval strategy is to retrieve all the *relevant* documents at the same time retrieving as few of the *non-relevant* as possible. When the characterisation of a document is worked out, it should be such that when the document it represents is relevant to a query, it will enable the document to be retrieved in response to that query. Human indexers have traditionally characterised documents in this way when assigning index terms to documents. The indexer attempts to anticipate the kind of index terms a user would employ to retrieve each document whose content he is about to describe. Implicity he is constructing queries for which the document is relevant. When the indexing is done automatically it is assumed that by pushing the text of a document or query through the same automatic analysis, the output will be a representation of the content, and if the document is relevant to the query, a computational procedure will show this.

Intellectually it is possible for a human to establish the relevance of a document to a query. For a computer to do this we need to construct a model within which relevance decisions can be quantified. It is interesting to note that most research in information retrieval can be shown to have been concerned with different aspects of such a model.

An information retrieval system

Let me illustrate by means of a black box what a typical IR system would look like. The diagram shows three components: input, processor and output. Such a trichotomy may seem a little trite, but the components constitute a convenient set of pegs upon which to hang a discussion.

Starting with the input side of things. The main problem here is to obtain a representation of each document and query suitable for a computer to use. Let me emphasise that most computer-based retrieval

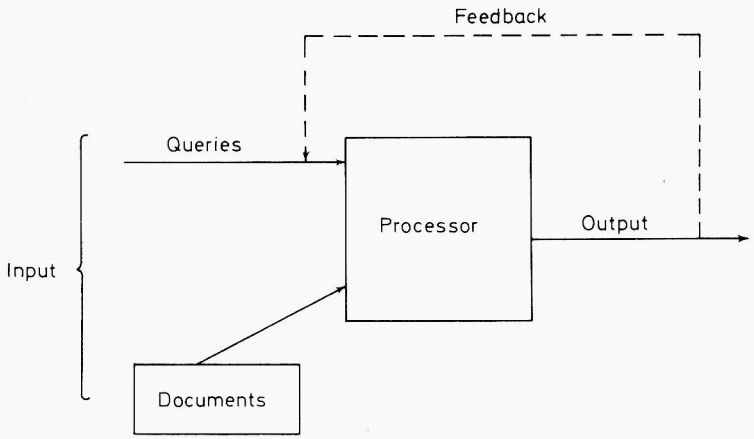

Figure 1.1. A typical IR system

systems store only a representation of the document (or query) which means that the text of a document is lost once it has been processed for the purpose of generating its representation. A *document representative* could, for example, be a list of extracted words considered to be significant. Rather than have the computer process the natural language, an alternative approach is to have an artificial language within which all queries and documents can be formulated. There is some evidence to show that this can be effective (Barber *et al.*[6]). Of course it presupposes that a user is willing to be taught to express his information need in the language.

When the retrieval system is on-line it is possible for the user to change his request during one search session in the light of a sample retrieval, thereby it is hoped improving the subsequent retrieval run. Such a procedure is commonly referred to as *feedback*. An example of a sophisticated on-line retrieval system is the MEDLINE system (McCarn and Leiter[7]). I think it is fair to say that it will be only a short time before all retrieval systems will be on-line.

Secondly, the processor, that part of the retrieval system concerned with the retrieval process. The process may involve structuring the information in some appropriate way, such as classifying it. It will also involve performing the actual retrieval function, that is, executing the search strategy in response to a query. In the diagram the documents have been placed in a separate box to emphasise the fact that they are not just input but can be used during the retrieval process in such a way that their structure is more correctly seen as part of the retrieval process.

Finally, we come to the output, which is usually a set of citations or document numbers. In an operational system the story ends here. However, in an experimental system it leaves the evaluation to be done.

IR in perspective

This section is not meant to constitute an attempt at an exhaustive and complete account of the historical development of IR. In any case it would not be able to improve on the accounts given by Cleverdon[8] and Salton[9]. Although information retrieval can be subdivided in many ways, it seems that there are three main areas of research which between them make up a considerable portion of the subject. They are: content analysis, information structures, and evaluation. Briefly the first is concerned with describing the contents of documents in a form suitable for computer processing; the second with exploiting relationships between documents to improve the efficiency and effectiveness of retrieval strategies; the third with the measurement of the effectiveness of retrieval.

Since the emphasis in this book is on a particular approach to document representation, I shall restrict myself here to a few remarks about its history. I am referring to the approach pioneered by Luhn[10]. He used frequency counts of words in the document text to determine which words were sufficiently significant to represent or characterise the document in the computer (more details about this in the next chapter). Thus a list of what might be called 'keywords' was derived for each document. In addition the frequency of occurrence of these words in the body of the text could also be used to indicate a degree of significance. This provided a simple weighting scheme for the 'keywords' in each list and made available a document representative in the form of a 'weighted keyword description'.

At this point it may be convenient to elaborate on the use of 'keyword'. It has become common practice in the IR literature to refer to descriptive items extracted from text as *keywords* or *terms*. Such items are often the outcome of some process such as, for example, the gathering together of different morphological variants of the same word. In this book keyword and term will be used interchangeably.

The use of statistical information about distributions of words in documents was further exploited by Maron and Kuhns[11] and Stiles[12] who obtained statistical associations between keywords. These associations provided a basis for the construction of a thesaurus as an aid to retrieval. Much of this early research was brought together with the publication of the 1964 Washington Symposium on *Statistical Association Methods for Mechanized Documentation* (Stevens *et al.*[13]).

Sparck Jones has carried on this work using measures of association between keywords based on their frequency of co-occurrence (that is, the frequency with which any two keywords occur together in the same document). She has shown[14] that such related words can be used effectively to improve recall, that is, to increase the proportion of the relevant documents which are retrieved. Interestingly, the early ideas of Luhn are still being developed and many automatic methods of characterisation are based on his early work.

The term information structure (for want of better words) covers specifically a logical organisation of information, such as document representatives, for the purpose of information retrieval. The development in information structures has been fairly recent. The main reason for the slowness of development in this area of information retrieval is that for a long time no one realised that computers would not give an acceptable retrieval time with a large document set unless some logical structure was imposed on it. In fact, owners of large data-bases are still loath to try out new organisation techniques promising faster and better retrieval. The slowness to recognise and adopt new techniques is mainly due to the scantiness of the experimental evidence backing them. The earlier experiments with document retrieval systems usually adopted a serial file organisation which, although it was efficient when a sufficiently large number of queries was processed simultaneously in a batch mode, proved inadequate if each query required a short real time response. The popular organisation to be adopted instead was the inverted file. By some this has been found to be restrictive (Salton[15]). More recently experiments have attempted to demonstrate the superiority of clustered files for on-line retrieval.

The organisation of these files is produced by an automatic classification method. Good[16] and Fairthorne[17] were among the first to suggest that automatic classification might prove useful in document retrieval. Not until several years later were serious experiments carried out in document clustering (Doyle[18]; Rocchio[19]). All experiments so far have been on a small scale. Since clustering only comes into its own when the scale is increased it is hoped that this book may encourage some large scale experiments by bringing together many of the necessary tools.

Evaluation of retrieval systems has proved extremely difficult. Senko[20] in an excellent survey paper states: 'Without a doubt system evaluation is the most troublesome area in ISR . . .', and I am inclined to agree. Despite excellent pioneering work done by Cleverdon *et al.*[21] in this area, and despite numerous measures of effectiveness that have been proposed (see Robertson[22, 23] for a substantial list), a general theory of evaluation had not emerged. I attempt to provide foundations for such a theory in Chapter 7 (page 168).