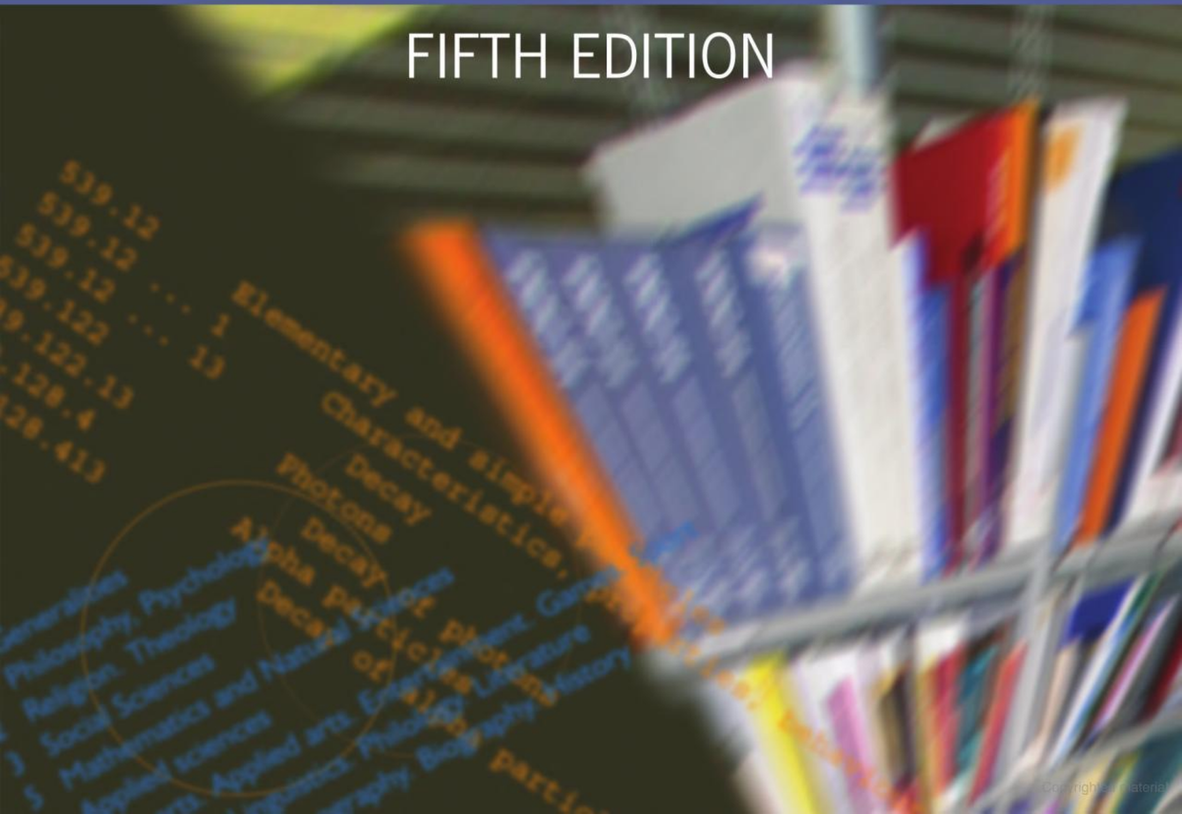
A C Foskett

The Subject Approach to Information

FIFTH EDITION



The Subject Approach to Information

Fifth edition

A C Foskett MA FLA AALIA
School of Communication and Information Studies,
University of South Australia

O What a tangled web we weave
When first we practise to retrieve . . .

facet publishing

Published by
Facet Publishing
7 Ridgmount Street
London WC1E 7AE

Facet Publishing is wholly owned by CILIP: the Chartered Institute of Library and Information Professionals

First published by Library Association Publishing 1969
Second edition 1971
Third edition 1977
Fourth edition 1982
This fifth edition 1996
Reprinted 1997
Reprinted by Facet Publishing 2005, 2007

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN 978-1-85604-048-8

Typeset in 10/12pt Times and Arial by Library Association Publishing.
Printed and made in Great Britain by MPG Digital, Bodmin, Cornwall.

Contents

Preface

'Another damn'd thick, square, book!', exclaimed the Duke of Gloucester, on being presented with the second volume of *The decline and fall of the Roman empire*. I do not aspire to the heights of that work, but I must admit that the description might be applied to the present volume; in order to cover my subject adequately I have had to range widely across the whole field of information retrieval. Attempts to cut down in any one area have been thwarted by the need to extend the coverage in others in order to achieve my objectives.

I produced the first edition in 1969 to meet what seemed to me to be a real need for a textbook which covered all aspects of the subject approach: not just classification or subject cataloguing (then taught as separate subjects), but also the possibilities being opened up by the computer. The Cranfield projects had shown that all information indexing languages are basically the same, and I had found that students taught these underlying principles could not only use existing classification schemes, lists of subject headings or thesauri, but could also construct their own in class assignments.

The first edition appears to have met a widespread need, and the second edition was published in 1971, incorporating revisions suggested by reviewers or arising from my own use of the book as a class text. This edition was also well received, and was translated into Portuguese for use in library schools in Brazil. In view of the continued demand, I intended to produce the third edition in 1976 as part of the centenary celebrations, but health problems made this impossible, and it appeared in 1977. The fourth edition was published in 1982, and its use became world-wide, including production as a talking book for use in Scandinavia. My original plan was to produce a new edition in 1987, but anyone who has been through the academic upheavals of the last decade will appreciate that my time was otherwise occupied!

When I did settle down to write this edition, my intention was to revise the 1982 edition, bringing descriptions of the various schemes up to date, but retaining most of the rest. It soon became apparent that the use of computers has made such major changes to the process of information retrieval that a far more radical revision was needed, and most of the book has in fact been rewritten. The basic concepts of recall and relevance remain the same, so Chapter 2 has not changed greatly; little recent work has been done on main classes in general classification schemes, so Chapter 10 has also changed little. Every other chapter has either been heavily revised or completely rewritten.

One area which has changed in every edition has been the treatment of the computer and its significance. In the first and second editions, the section on 'the future' had a chapter on the computer. By the third edition, the ‘future’ had become the present, and I covered the computer in Chapter 3, just before derived indexing. By 1982, it was clear that those using the text – mainly library school students – would have covered computing in sufficient detail elsewhere in their courses for me to be able to take it for granted, and refer to it as necessary. For this edition, I have in a sense gone back a step by including a chapter on information technology. My purpose is to attempt to show how developments over the past 30 years have radically changed the way we can now use computers: why, for example, it is usually easy to incorporate modern material into a computer-based system, but older materials present often costly problems. Those familiar with this background can easily GOTO Chapter 5.

Do we need to understand the theory of information retrieval any longer, now that so much material is available through the computer? I believe that the answer is a resounding YES; indeed, I believe that we must have a very clear idea of what we are doing if we are not to get lost in the mass of information at our fingertips. More serious is the risk that we run of being misled by the computer. The literature on online searching is full of examples of failures, where users have tried searching but found nothing. Even more dangerous is to find *something* and assume that what has been found is all that there is to be found. Also, recent developments have meant that we have a wider choice of search strategies; tools such as DDC, UDC and LCSH are now available online, and LCC is being processed, so that we can use a variety of approaches in online searching which were not available until quite recently – but we must know how to use them. CD-ROM, the Internet, and the World Wide Web have made vast quantities of information accessible to us – if we know where to find it, and how to evaluate it. The fact that information is available on computer does not mean that it is necessarily better than that available in print; indeed, most material in print goes through a process of evaluation which is noticeably absent from most Internet publication. I have tried to cover the methods used in online retrieval in Chapter 5, and I have returned to the question of control and evaluation in Chapter 28, for it is in this sphere that I see much of the future role of the information worker.

Previous editions have not specifically dealt with thesaurus construction, though it is implicit in the discussion of semantic relationships in Chapter 6. In view of the increased significance of thesauri I have made it explicit in this edition, showing how the analysis needed to establish relationships between terms in a subject area is in fact the basis of a thesaurus. A discussion of syntactic relationships in pre-coordinate indexing follows, leading into a chapter on alphabetical headings (including PRECIS), and five chapters on various aspects of systematic arrangement: classification. This does not mean that systematic arrangement is more important than alphabetical headings, but simply that there is more to discuss. PRECIS rated a chapter to itself in the fourth edition; as its use has been discontinued by BNB, it could perhaps have been relegated to history; however, it seemed useful to have a reasonably full description of the system as set out in the second edition of the *Manual*, which includes several features not in the earlier version.

Why do we continue to use pre-coordinate indexing despite its disadvantages?

Chapter 14 looks at manual pre-coordinate indexes, including the card and fiche catalogues which are still to be found in many libraries, while Chapter 15 is a completely new look at OPACs and the MARC records which have made them possible. Chapters 16 to 24 look at the widely used classification schemes and lists of subject headings, including a new chapter on the Broad System of Ordering. In such a review, chapters become superseded as new editions are published; in this instance, a new edition of DDC will be published very shortly after the appearance of this book. Had the book been published at any other time, some other scheme would certainly have appeared in a new edition shortly afterwards. This is a battle the author cannot win! However, I have been able to include information about DDC21, and nearly all the chapter will remain relevant in any case.

There seemed to be little point in including a chapter on manual methods of post-coordinate indexing; though these were still widely used in 1982, by 1996 they have become museum pieces. Anyone who needs a description of edge-notched, Uniterm or peek-a-boo cards can go back to an earlier edition. Similarly, the chapter on computer-based systems has been overtaken by events. There are now too many to describe in detail, while the principles on which they work are covered in Chapter 5. Videotex and teletext are not covered; they appear to have fallen by the wayside as significant providers of information.

In addition to the chapters on thesauri in science and technology, and the social sciences, I have added one on visual arts and graphics. The use of graphics on computers is now so significant that it seemed appropriate to look at two of the schemes which have been developed to control this kind of information. Both the classification scheme *Iconclass* and the *Art and Architecture Thesaurus* are of interest in themselves, as well as being the means of keeping control of the information available in this format.

The results of the Cranfield projects and those that followed have now been assimilated to such an extent into retrieval theory that there was no longer any point in having a chapter on the evaluation of IR systems; in its place I have reverted to the practice of the early editions by concluding with a chapter on ‘The future’. In a field such as this, to write about the future is tempting fate, and I know that ‘the digital/virtual/electronic library’ already exists; however, I feel that we still have some way to go before it becomes commonplace. Many of the problems to be solved are not technological but legal and ethical: intellectual property and integrity of communication are two that present real difficulties. In addition to these aspects of the digital library, I have tried to summarize the situation in medicine as a case study of possible future developments in other subjects. The Cochrane Collaboration is an excellent example of how one kind of literature can be controlled and exploited to good advantage. It is also of course very much in line with the need for more review articles expressed at the Royal Society’s Scientific Information Conference – but that was in 1948! Surely anything that old is irrelevant to the future of information retrieval?

As in previous editions, I do not present any basic philosophical arguments for my ideas. They are based on a behavioural approach: what do we have to *do* to retrieve information? If our first attempts are unsuccessful, how can we amend them, and what kinds of tool are likely to help us? It seems to me that the whole of information retrieval stems from those very simple questions, without the need to look for a philosophy. I have tried, as always, to make the presentation *readable*; I see no reason why textbooks should be boring. From time to time, I have slipped in a remark which is perhaps not to be taken too seriously. I recently had a comment from a librarian who had used the fourth edition as a student, that spotting the jokes was one of the things that made using my text popular!

To assist users, I have put words which I think are significant in italics; these are usually terms which are defined, or are important in the context. They are the sort of word which might be selected for the heading of a note by a student, or used in a handout or overhead by a lecturer. Italics are also used for titles of books and periodicals, in line with usual conventions. In describing schemes in Chapters 16 to 27, I have tried to copy the text as closely as possible for the examples I give, so that the various combinations of italic, bold, roman, sans serif, caps, small caps and lower case, and indentation are as they appear in the originals. However, I must stress that I quote *examples*, and I do so selectively to show the points I am trying to make. Students need to look at the originals at first hand to gain a full appreciation of the scheme being described. My purpose is to make such first hand study informed and thus profitable, not to replace it.

Acknowledgments

Any work such as this owes a great deal to help received from colleagues. In particular, I am indebted to Dr Ia McIlwaine of UCL for her very full and helpful comments, first on my outline and later on the first draft; without these, the book would have been a great deal less effective than I hope it now is. For information on specific chapters I have to thank the following: Chapters 4 and 5, my colleague Sue Myburgh; DDC, Joan Mitchell (Editor), Ross Trotter and Giles Martin; UDC, Ia McIlwaine (Editor, English Medium edition); BC, Jack Mills; BSO, Eric Coates; LCSH, my colleague Derek Wiblin; Sears, Joseph Miller (Editor); ROOT Thesaurus, Jean Aitchison; INSPEC Thesaurus and Classification, Jeff Pache; ERIC Thesaurus, Jim Houston (Lexicographer); the Cochrane Collaboration, Michael Brittain. For the many others who have helped form my thinking over the past years, I hope that the appearance of their names in the chapter bibliographies will prove an adequate recognition. As always, I must conclude by thanking my Wife for her continuing support over the years, without which neither this nor any of the previous editions would have appeared.

A C Foskett
February 1996

List of abbreviations

AAAS	American Association for the Advancement of Science
AACR	Anglo-American Cataloguing Rules
AAL	Association of Assistant Librarians
AARNET	Australian Academic Research NETwork
AAT	Art and Architecture Thesaurus
ABN	Australian Bibliographic Network
ABNO	All But Not Only
ACM	Association for Computing Machinery
ACS	American Chemical Society
ADDC	Abridged Dewey Decimal Classification
ADFA	Australian Defence Force Academy
AFP	Anticipated Futility Point
AID	Associative Interactive Dictionary
AIP/UDC	American Institute of Physics/Universal Decimal Classification
ALA	American Library Association
ANB	Australian National Bibliography
ANSEL	Extended Latin character set
ANSI	American National Standards Institute
APUPA	Alien-Penumbra-Umbra-Penumbra-Alien
ARPANET	Advanced Research Projects Agency NETwork
ASCA	Automatic Subject Citation Alert
ASCII	American Standard Code for Information Interchange
ASCIS	Australian Schools Cataloguing Information Service
ASSIA	Applied Social Science Index and Abstracts
ASTIA	Armed Services Technical Information Agency
BBS	Bulletin Board Service
BC	Bliss Classification/ Bibliographic Classification
BCA	Bliss Classification Association
BL	British Library
BLAISE	British Library Automated Information SErvice
BLBSD	British Library Bibliographical Services Division
BLR&DD	British Library Research & Development Department
BNB	British National Bibliography
BSI	British Standards Institution
BSO	Broad System of Ordering
BTI	British Technology Index
CATLINE	Current CATalogue onLINE (National Library of Medicine)

CATNI	Catchword And Trade Name Index
CC	Colon Classification
CC	Current Contents (ISI)
CCC	Central Classification Committee (UDC)
CCF	Common Communications Format
CCML	Comprehensive Core Medical Library
CD	Compact Disk
CD-ROM	Compact Disk-Read Only Memory
CDS	Cataloguing Distribution Service
CERN	Centre for High-Energy Physics
CIA	Central Intelligence Agency
CIJE	Current Index to Journals in Education
CIM	Cumulated Index Medicus
CITE NLM	Current Information Transfer in English, National Library of Medicine
COM	Computer Output Microform/fiche/film
COMPASS	COMPuter Aided Subject System
CORE	Chemistry Online Retrieval Experiment
COSATI	Committee on Scientific And Technical Information
CRG	Classification Research Group
CTI	Current Technology Index
DC&	Decimal Classification: Additions, Notes, Decisions
DDC	Dewey Decimal Classification
DLA	Division of Library Automation (University of California)
DoD	Department of Defense
EdNA	Australian Education Network
EE	English Electric
EJC	Engineers Joint Council
EMPST	Energy-Matter-Personality-Space-Time
ERIC	Educational Resources Information Clearinghouse
ESP	Extended Subject Program
ESS	Editorial Support System (DDC)
FID	Fédération Internationale d'Information et de Documentation
FP	Futility Point
FTP	File Transfer Protocol
GARE	Guidelines for Authorities and Reference Entries
GIF	Graphics Interchange Format
GUI	Graphical User Interface
HTML	HyperText Markup Language
HTTP	HyperText Transport Protocol
IAIMS	Integrated Academic Information Management System
IBM	International Business Machines
ICI	Imperial Chemical Industries Ltd.
ICSU	International Council of Scientific Unions
ICT	International Critical Tables

IEE	Institution of Electrical Engineers
IEEE	Institute of Electrical and Electronic Engineers
IFLA	International Federation of Library Associations and Organizations
IIB	Institut International de la Bibliographie
IID	Institut International de Documentation
IM	Index Medicus
IM	International MARC [now part of UBCIM]
INSPEC	INformation Service in Physics, Electrotechnology, Computers and control
IR	Information Retrieval
ISBD	International Standard Bibliographical Description
ISDN	Integrated Services Digital Network
ISI	Institute for Scientific Information
ISILT	Information Science Index Languages Test
ISO	International Organization for Standardization
ISONET	ISO Network
IT	Information Technology
JANET	Joint Academic NETwork
JPEG	Joint Picture Experts Group
KWIC	KeyWord In Context
KWOC	KeyWord Out of Context
LAN	Local Area Network
LASER	London And South-Eastern Library Region
LC	Library of Congress
LCC	Library of Congress Classification
LCCN	Library of Congress Control Number
LCSH	Library of Congress Subject Headings
LISA	Library and Information Science Abstracts
LP	Long Play (records)
LUCIS	London University Computer Information Service
MARC	MAchine Readable Cataloguing
MEDLARS	MEDical Literature Analysis and Retrieval System
MEDLINE	MEDLARS onLINE
MELVYL	OPAC used in the University of California libraries
MeSH	Medical Subject Headings
MIT	Massachusetts Institute of Technology
MRF	Master Reference File (UDC)
NAL	National Agricultural Library
NASA	National Aeronautics and Space Administration
NATO	North Atlantic Treaty Organization
NBS	National Bibliographic Service
NCSA	National Center for Supercomputer Applications
NEH	National Endowment for the Humanities
NEPHIS	NEsted PHrase Indexing System

NEXUS	Schools network, South Australia
NISO	National Information Standards Organization
NLM	National Library of Medicine
NREN	National Research and Education Network
NSF	National Science Foundation
NSFNET	National Science Foundation NETwork
NTIS	National Technical Information Service
OBNA	Only But Not All
OCLC	Online Computer Library Center
ODA	Office Document Architecture
ODIF	Office Document Interchange Format
OKAPI	Online Keyword Access to Public Information
OPAC	Online Public Access Catalogue
OSI	Open Systems Interconnection
OSTI	Office for Scientific and Technical Information (later BLR&DD)
PAIS	Public Affairs Information Service
PCL	Polytechnic of Central London
PMEST	Personality-Matter-Energy-Space-Time
PRECIS	PREserved Context Indexing System
RIE	Resources In Education
RIN	Reference Indicator Number
RMIT	Royal Melbourne Institute of Technology
SAERIS	South Australian Educational Resources Information Service
SAMOS	Satellite And Missile Observation System
SCI	Science Citation Index
SCIS	Schools Cataloguing Information Service
SDI	Selective Dissemination of Information
SDIF	SGML Document Interchange Format
SGML	Standardized General Markup Language
SIN	Subject Indicator Number
SMART	Experimental computer-based IR system devised by G. Salton
SPDL	Standard Page Description Language
SRC	Standard Reference Code/Standard Roof Classification
SRIS	Science Reference and Information Service
STAIRS	STorage And Information Retrieval System (IBM)
TCP/IP	Transmission Control Protocol/Internet Protocol
TEI	Text Encoding Initiative
TEST	Thesaurus of Engineering and Scientific Terms
UBC	Universal Bibliographical Control
UBCIM	Universal Bibliographical Control – International MARC
UC	University of California
UDC	Universal Decimal Classification
UDCC	Universal Decimal Classification Consortium
UKAEA	United Kingdom Atomic Energy Authority

UKOLN	United Kingdom Online Library Network
UMLS	Unified Medical Language System
UNIMARC	Universal MARC format
UNISIST	World science information system
URL	Uniform/Universal Resource Locator
USAEC	United States Atomic Energy Commission
VINITI	All-Union Institute for Scientific and Technical Information
VIP	Vocabulary Improvement Project (ERIC)
WAN	Wide Area Network
WASP	White Anglo-Saxon Protestant
WLN	Washington Library Network
WWW	World Wide Web

Part I

Theory of information retrieval systems

Chapter 1

Introduction

It is frequently said that we live in the 'Information Age', and nearly every day we learn of some new development in information technology. The human need for information is growing, as our societies grow to depend more and more on information to survive and flourish. We all need to be able to find facts, but we also need to find *information* on particular subjects – not just the bare facts, but their evaluation and assimilation into our own frame of reference. Can the computer solve all our problems, or do we still need to bring human intelligence to bear on the solutions? How do we set about finding information that we need? This book attempts to look at the kind of problems we meet in trying to find information that meets our needs, how the computer can help, and how human effort can still be valuable in easing our path to discovery.

It is helpful if we try to define some terms so that we can see their relevance to our daily lives, and to the work of the professional information worker. The following definitions are based on those in the *Concise Oxford dictionary*[1] and the *Macquarie Dictionary*.[2]

- *knowledge* is what *I* know
- *information* is what *we* know, i.e. *shared* knowledge
- *communication* is the imparting or interchange of . . . information by speech, writing or signs, i.e. the *transfer* of information
- *data* [literally things given] any fact(s) assumed to be a matter of direct observation.
- Additionally, a *document* is any physical form of recorded information.

From these definitions we can see that data consists of unprocessed facts; knowledge is what an individual possesses after assimilating facts and putting them into context; information is knowledge shared by having been communicated. *Information technology* is the equipment, hardware and software that enables us to store and communicate large amounts of data at high speed. If we record knowledge, then it may be communicated at a distance in space and time; we do not have to be face to face with the informant as we do in oral communication. This further suggests the concept of a *repository* or store of recorded information.

Before knowledge was recorded, individuals formed the repository of knowledge, the bridge between successive generations and between those who generated new information and those who required to use it. The amount of information that can be passed on in this way is limited, and society began to move forward when information of various kinds began to be recorded in relatively permanent forms which could serve as a substitute for the 'elder' in person. Knowledge only becomes generally useful when it is communicated; by recording it, we do our best to ensure that it is permanently available to anyone who may need it, instead of ephemeral and limited to one individual.

Nowadays, the amount of new information being generated is such that no individual can hope to keep pace with even a small fraction of it, and the problem that we have to face is that of ensuring that individuals who need information can obtain it with the minimum of cost (both in time and in money), and without being overwhelmed by large amounts of irrelevant matter. Sherlock Holmes[3] puts the matter well:

> . . . a man should keep his little brain attic stocked with all the furniture he is likely to use, and the rest he can put away in the lumber-room of his library, where he can get it if he wants it.

Holmes himself kept not only a library of published works, but also his own personal index, to which he referred on many occasions to supplement his own knowledge. The point is that we do not have to know everything – but we must know how to find information when we need it.

So, instead of the individual store of knowledge, we now have the corporate store: libraries, information services, computers. Instead of the individual memory, we have the corporate memory: library catalogues, bibliographies, computer databases. And just as the individual whose memory fails cannot pass on wanted information, so the inadequate corporate memory will fail in its purpose. We have to ensure that the tools we prepare meet the needs of our users. It is therefore very important to try to define the needs of our users as closely as possible, particularly in view of the exponential growth of knowledge in recent years. Professional librarianship is concerned with the skills, both human and technical, needed to plan and use systems which will achieve the optimum results in meeting the needs of users. Libraries are still the main repositories of information, but computers have now become the most significant factor among the tools used.

The growth of knowledge

A valuable study[4] identified three eras of information need, to which we may add a preliminary fourth, the polymath era. There was a time when the sum total of human knowledge was sufficiently small to be comprehended by one individual. As knowledge grew, we moved into the discipline-oriented era, which lasted in effect from the invention of printing until well into the twentieth century. This was characterized by the division of knowledge into more or less water-tight compartments or *disciplines*, reflecting the way in which they were studied; new disciplines grew out of the splitting up, or 'fission' of existing disciplines as particular aspects grew in importance and developed into disciplines in their own right. Thus science developed from philosophy as a field of study; physics developed from science; electricity developed from physics; and electronics developed from electricity. In each case the new subject represented a fragmentation of the old, but remained within it. Most of the conventional retrieval tools used in libraries were developed within this frame of reference, and it is only as new eras, with changing user needs, have developed that schemes such as DDC and other library tools have begun to show serious signs of strain. We shall be examining some of these problems later.

The second era, the problem-oriented era, began to assume importance in the 1930s, and particularly in the Second World War. This was characterized by the need to solve particular problems, using whatever disciplines might be necessary, regardless of whether they 'belonged together' or not. A recent example comes from the field of micro-engineering, where parts are so small that novel methods of moving them have to be devised. Japanese scientists have 'borrowed' the methods used by cilia (the fine hairs found on mucous membranes) to move mucus, in order to be able to move the microscopic units involved. The best-known example is genetic engineering, which involves the merging, or 'fusion', of disciplines such as physics and biology which used to be thought of as fundamentally separate.

We are now in the third, or mission-oriented, era, in which demands for information may span a range of disciplines. For example, space medicine certainly requires a knowledge of medicine, but in addition involves problems related to space physics, mechanics (the phenomenon of weightlessness), diet, hygiene – the list is formidable. Clearly the old barriers between disciplines, which began to crumble in the problem-oriented era, have now effectively disappeared, presenting further difficulties in the communication of information. The more remote new information is from individuals' existing range of knowledge, the more difficult it becomes for them to comprehend and incorporate it into their own store of knowledge. The needs of today's information users place demands on information services far more acute than those experienced in past eras, and our information retrieval systems must be adequately developed to meet these demands.

It is also useful to distinguish between the retrieval of data and the retrieval of information. Retrieving information from documents is not the same as retrieving data – there are some important differences.[5] A request for data is satisfied by directly providing the desired fact(s); a request for information is satisfied by providing either references to documents or the documents themselves which will *probably* contain the desired information. Requests for data are deterministic and require no logical decisions on the part of the enquirer; the answer supplied is either right or wrong. On the other hand, requests for information are probabilistic, and may involve a series of logical decisions on the part of the enquirer. So a request for data should lead to the *right* answer, otherwise it is useless. By contrast, a request for information should lead to a *useful* answer, which does not necessarily have to be complete; its usefulness is a matter of judgement on the part of the enquirer.

Information retrieval as a form of communication

In the light of the preceding discussion, we may see that one measure of the success of an information retrieval system is its effectiveness as a means of communicating information. It is therefore helpful to look at the communication process itself, and the ways in which it is modified by the information retrieval process.

We may consider information retrieval processes as part of the overall pattern of communication. The most commonly used model of the communication process is that devised by Shannon and Weaver, shown in Figure 1.1(a). In this model, we see that a *source* has a *message* which is to be transmitted to a *receiver*; before it can be transmitted, the message must be *encoded* for transmission along the selected *channel*, to be *decoded* before it can be understood by the receiver. In information retrieval, the sources are the originators of the documents we handle; the encoding process includes the choice of the appropriate physical manifestations – words, sounds, images – and their translation into an appropriate medium; the channel is

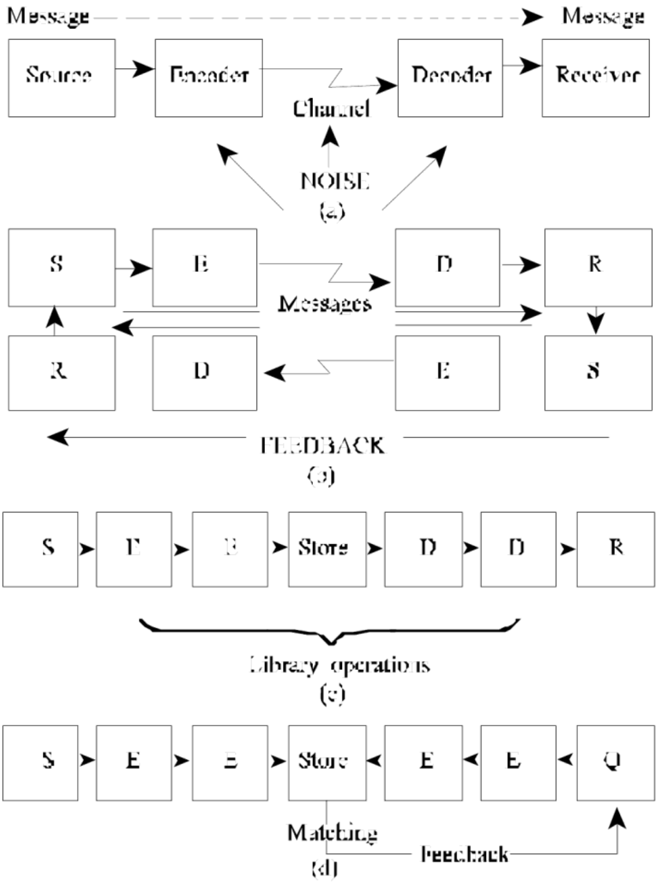


Fig. 1.1 Models of the communication process
(a) The Shannon–Weaver model
(b) Verbal (two-way) communication; involves feedback
(c) The effect of library operations
(d) The query situation: feedback through the matching process

the resulting document and its progress from source to receiver; and the decoding process involves the receiver's ability to comprehend the message in the form in which it is presented. The final element in the model is *noise*. Noise may be defined as anything which detracts from the fidelity of transmission of the message from source to receiver. Shannon and Weaver were concerned with the transmission of messages over telephone wires, but the concept of noise can be generalized to cover all kinds of interference with communication – for example, the retrieval of unwanted documents in response to a request.

If we consider normal verbal communication, we can see that the model (Figure 1.1(b)) is in fact that of Figure 1.1(a) doubled, so that the original source functions also as a receiver, and vice versa. In this situation, a further important element enters the picture: the idea of *feedback*. If the message becomes distorted on its way from source to receiver, the receiver can immediately query it (I didn't quite catch that, or, Could you explain that again, please?, or, *What*?) Feedback can thus be significant in reducing the effects of noise.

Unfortunately, when we are dealing with documents, there can be little, if any, feedback from receiver to source. The source has no control over who receives messages, and cannot therefore restrict them to a specific audience. Receivers in turn cannot be sure that they have understood a message correctly, or that they have located all – or indeed any – of the messages they were looking for. Indeed, the situation is made more difficult by the interposition of additional encoding processes, which then require further decoding processes on the part of the receiver. In libraries, we normally arrange books by putting a code on the spine: a call number; and we identify them in our catalogues by a further series of codes: catalogue entries (Figure 1.1(c)). Catalogue entries and class marks can in fact give a great deal of information, but for most library users they form an additional complication in the chain of communication, and are an inherent additional source of noise. Furthermore, we have introduced another complication: transmission of messages is delayed by their being placed in a *store* of some kind. A book may be regarded as such a store, but so may a CD-ROM encyclopedia or any other kind of document; libraries are stores, as are bibliographic databases. In this context, relatively trivial matters such as the misshelving of a book or a power supply malfunction all add their quota of noise to the communication process.

The importance of bibliographic control

Documents can take a variety of physical forms. If we consider a library as a convenient example of a store of documents, at one time we could have assumed that for many libraries the bulk of the stock of documents would have been books; in others, periodicals or technical reports might predominate. Now, libraries may contain audiovisual and computer-readable materials as well as printed, and the computer-readable materials may themselves contain the equivalent of printed and audiovisual materials in the form of multimedia presentations. We may then have the same work in several different physical forms. We may have Shakespeare's play *Hamlet* as a printed book, or as a film, or spoken word cassette. An encyclopedia may be a set of printed volumes or a single CD-ROM. The *intellectual content* remains the same, but the physical format varies.[6] It is the *work* which is important, but we can no longer rely on the physical arrangement of the documents to bring together different versions of the same work. We have to rely on a substitute – a set of surrogate *records* – of the documents we have, in order to achieve *bibliographic control*. Hagler has summarized the purposes of bibliographical control:[7]

1 identifying the existence of all possible documents produced in every physical medium;
2 Identifying the works contained in documents or as parts of them (e.g. periodical articles, conference papers, contents of anthologies etc.);
3 producing lists of these documents and works prepared according to standard rules of citation;
4 providing all useful access points (indexes) to these lists, including at least some access by name, title and subject;
5 providing some means of locating a copy of each document in a library or other accessible collection.

These records must use *words*; we can only identify and locate information if we can adequately describe it in words. (Even in a collection of art materials, where it is possible to match, say, specific examples of textures with illustrations in which they occur, it is still necessary to identify *textures* as the key word in our search.)

The situation has been further complicated by a historical division between the bibliographic control of books and that of other materials. Books have normally been treated as units, and listed – catalogued – as such, whereas since the development of the scientific periodical in the seventeenth century, and in other fields later, periodicals have been controlled through a quite separate mechanism of abstracts and indexes, normally not produced by librarians. With the development of audiovisual materials, a new apparatus of bibliographic control grew up to enable users to find what they needed. These divisions have been to some extent overcome by the development of computer databases, but there is still a major division between the bibliographic control of books and that of non-book materials. Even online library catalogues – OPACS, discussed in Chapter 15, do not normally cross this divide; for the bibliographic control of other materials we turn to other databases – though to confuse the issue still further these may be available through an OPAC.

If we only have one kind of document in our collection, for example a set of records in a computer database, we still have to have access to the records from various points of view. We might wish to identify all those on a particular subject, or by a particular author, or with a particular title, as indicated in Hagler's point 4 above. If the database contains images, we may wish to locate all the records containing a particular image or kind of image. To find specific items, bibliographic control is still needed to give us access to them through a variety of factors.

Some of these factors *identify* the item(s) they refer to. For example, if we have the number of a patent specification, there will be only one item corresponding to that description. If we have selected a particular author, then the number of works which will match our requirements is immediately limited. If we choose a particular title, we shall find only one item, or perhaps a few, which will satisfy our request (authors usually try to find unique titles for their works, but do not always display enough originality!). Furthermore, there are now standards which help us to determine the format in which we should look for the information we need; for example, the widely used Anglo-American Cataloguing Rules[8] and the MARC records (discussed in Chapter 15) now enable us to formulate a search for a factor that identifies in such a way that it becomes a search for data rather than for information.

Factors which do not identify

If we are asked for information on a particular subject, we face a rather different set of problems. To begin with, there is no set of standards to tell us precisely how to express the subject, and indeed virtually every database seems to have its own rules. Readers seek information on particular subjects, and expect our systems to be able to provide the answers. In this situation, the readers/receivers become sources, encoding messages in the form of enquiries. (Figure 1.1(d)). We now have to discover any messages in our store which appear to match these enquiries; having found some, we can pass them on to the enquirers, who can decide whether the answers match their needs. In the light of our responses, enquirers may modify their messages in an attempt to achieve a closer match with their requirements; in other words, we have a degree of feedback in the system, which may enable us eventually to satisfy a request despite initial failure. This failure may arise from a variety of causes: enquirers may not be able to express their needs clearly, or may not be very sure exactly what those needs are (if they knew the answers they would not need to ask the questions!); or our encoding processes may be inadequate; or the original sources (the authors) may not have made their messages clear, or may even have had some rather different messages in mind. For example, the answer to a question from a librarian on the optimum number of people required to staff the circulation desk may be found in a book on supermarket management. Authors write within their own particular frame of reference, which will not be the same as that of the readers. We have to try to optimize the results of any search, while accepting that we can no longer characterize the result as right or wrong.

Reference retrieval, document retrieval and information retrieval

It is also important to recognize the distinction between reference retrieval, document retrieval and information retrieval. Surrogate records used for bibliographic control, such as conventional library catalogues and bibliographies, give us *reference* retrieval; we still have to supply the actual documents. The user has then to look at the documents, and can only make a final judgment on the basis of the information in the documents. Some computer-based systems, for example those in the legal field, have included full text for some years, and thus give document retrieval at the same time as reference retrieval, but at present the majority do not, though this is rapidly changing with the development of CD-ROM and multimedia. Further, if the communication process between authors and user is to be complete, the librarian may have to act as intermediary and interpret the documents found by translating them into language which can be understood by the user. This may mean literally taking a text in, say, German and producing an English translation, or it may involve an explanation process within the one language. In either case, it is a process which is often neglected; all too often the list of references is seen as the end product, not the actual communication of information.

This book is concerned with a discussion of the problems of optimizing our responses to requests for information on subjects. This is not to suggest that identifying factors such as authors' names do not present any problems; the fact that it took some 20 years of discussions to produce a new edition of the Anglo-American Rules, which has since been revised,[8] and is still the subject of discussion, shows very plainly that they do! The problems of the subject approach to information, however, are more severe because they are more indeterminate; we never reach the stage of being able to say we have finished a search conclusively. A great deal of research has been done on these problems; much more remains to be done. This book is an attempt to show the present state of the art in a way that will be acceptable as an elementary textbook; it does not pretend to be an advanced study, of which there are many,[9] but rather to give beginners some understanding of present theories and ideas.

References

1 *The concise Oxford dictionary*, 6th edn, Oxford, Clarendon Press, 1976.

2 *The Macquarie dictionary*, St Leonards, NSW, Macquarie Library Pty Ltd, 1981.

3 Doyle, Sir A. C., 'The adventure of the five orange pips', *The adventures of Sherlock Holmes*, 1892.

4 Arthur D. Little Inc, *Into the information age: a perspective for federal action on information*, Chicago, American Library Association, 1978.
The information society: issues and answers, edited by E. J. Losey, Phoenix, Oryx Press, 1978.

5 Blair, D. C., *Language and representation in information retrieval*, New York NY, Elsevier Science Publishers, 1990.

6 Hagler, R., *The bibliographic record and information technology*, 2nd edn, Chicago, USA, American Library Association; Ottawa, Canada, Canadian Library Association, c1991.

7 Hagler, R., ref. 6 above, 7.

8 *Anglo-American cataloguing rules*, 2nd edn, 1988 revision, London, Library Association, 1988.

9 Of the making of many books on information retrieval there appears to be no end. The following very select list is intended to act only as a guide to the student. Much of the basic work was published in the period 1960–1980 in such works as the Butterworth's series *Classification and indexing in . . .: science and technology*, by B. C. Vickery, 3rd edn, 1975; *the social sciences*, by D. J. Foskett, 2nd edn, 1975; *the humanities*, by D. Langridge, 1976. Specific refer-

ences will be made to these works in particular chapters, and a more complete list is included in the 4th edition of this work. The following list includes some of the works published since the previous edition.

Austin, D., *PRECIS: a manual of concept analysis and subject indexing*, 2nd edn, London, British Library, 1984.

Classification of library materials : current and future potential for providing access, Bengtson, B. G. and Hill, J. S. (eds.), Neal-Schuman, c1990.

Coates, E. J., *Subject catalogues: headings and structure*, reissued with new preface, London, Library Association, 1988.

Craven, T. C., *String indexing*, Orlando, Academic Press, 1986.

Hunter, E. J., *Classification made simple*, Aldershot, England; Brookfield, USA, Gower, c1988.

Lancaster, F. W. and Warner, A. J., *Information retrieval today*, Arlington, VA, Information Resources Press, 1993.

Langridge, D., *Subject analysis: principles and procedures*, London; New York, Bowker-Saur, 1989.

Milstead, J. L., *Subject access systems: alternatives in design*, Orlando, Academic Press, 1984.

Rowley, J. E., *Organising knowledge: an introduction to information retrieval*, 2nd edn, Aldershot, Ashgate, 1992.

Salton, G. and McGill, M. J., *Introduction to modern information retrieval*, New York, McGraw-Hill, c1983.

Subject access: report of a meeting sponsored by the Council on Library Resources Inc, Dublin, Ohio, 1982.

Dym, E. D. (ed.), *Subject and information analysis*, New York, M. Dekker, c1985.

Berman, S. (ed.), *Subject cataloging: critiques and innovations*, New York, Haworth Press, c1984.

Subject indexing: principles and practices in the 90's: IFLA satellite meeting August 17–18 1993, Lisbon, Munich, Saur, 1995.

Turner, C., *The basics of organizing information*, London, Bingley, 1985.

Chan, L.M., Richmond, P. A., Svenonius, E. (eds.), *Theory of subject analysis: a sourcebook*, Littleton, CO, Libraries Unlimited, 1985. A valuable collection of significant articles, referred to elsewhere as *Theory of subject analysis* . . .

Wynar, B. S. *Introduction to cataloging and classification*, 7th edn, by A. G. Taylor, Littleton, CO, Libraries Unlimited, 1985.

For developments over the past few years it is helpful to consult the chapters on classification in *British librarianship and information work*, London, Library Association, 1982–; 1976–1980, Taylor, L. J. (ed.), 1983; 1981–1985, Bromley, D. W. and Allott, A. M. (eds.), 1988; 1986–1990, Bromley, D. W. and Allott, A. M. (eds.), 1992; these also have comprehensive bibliographies. Useful series of articles include 'Subject access literature' annually in *Library resources and technical services*. Relevant chapters in the *Annual review of information science and technology* are also valuable for the student wishing to pursue the subject in depth.

Chapter 2

Features of an information retrieval system

Authors generate large quantities of information every day. Estimates made 30 or more years ago suggested that the number of useful (i.e. not merely repetitive) periodical articles published each year in science and technology alone was in excess of one million,[1] and the number has certainly increased since then. In Britain alone over 50,000 books are published each year, and the USA has now overtaken Britain as the world's most prolific book publisher. Libraries acquire a selection of this enormous output for the immediate use of their readers, and through the various schemes of interlibrary cooperation they have access to a very much wider choice. At the other end of the chain of communication we have readers, each with their own individual needs for information which has to be selected from the mass available. The readers' approach may be purposive, that is, they may be seeking the answer to specific questions, which may be more or less clearly formulated in their minds. This is the situation that we shall consider first, but we must not overlook the browsers, who are looking for something to catch their interest rather than answers to specific questions, and who form the majority of users in public libraries.

Information retrieval and document retrieval

We should distinguish between information – knowledge which is being communicated – and the physical means by which this communication takes place, as was pointed out in Chapter 1. In the past, although it has been the practice to refer to *information* retrieval, what has been described has been *document* retrieval; in other words, when asked for information, we have provided a set of documents which we believed would contain the information sought. The success of our search has been considered to be a subjective judgment, which could only be made by the individual making the request; indeed, many librarians have thought it beyond their terms of reference to make any attempt to *evaluate* the documents found unless the information sought was purely factual – and some not even then! This reticence has not been restricted to librarians; for example, the *International critical tables*, published 1927–1933, gave a selection of values for most physical data, indicating what was usually thought to be the 'best value', but giving chapter and verse for each value recorded so that users could make up their own minds if they wished. The US Academy of Sciences still maintains its Office of Critical Tables, though the sheer quantity of data now available has precluded publication of a revised edition of *ICT*.

Various attempts have been made to mechanize the document delivery part of the system, but none of these found widespread acceptance, mainly because of costs and mechanical problems. However, the development of new means of storing massive amounts of information in computers has led to a new approach to the direct retrieval of information. Factual information on a wide variety of subjects is now available in many countries through the various forms of videotext, using the TV set which is now an essential part of every civilized home (as George Orwell[2] pointed out). CD-ROM and videodisk provide access to equally large amounts of information, including sound and illustrations as well as text (multimedia), using a desktop computer. Networking allows information to be transmitted direct from computer to computer, making possible the paperless office discussed in Chapter 4 – though so far offices seem to be using even more paper than they did in the past. It has been suggested that paper consumption in the USA is likely to increase steadily until at least the year 2000!

We should not, and cannot, ignore these developments, but it remains important to recognize that they do not in fact alter the fundamentals of the communication process, though they may change some of its practical manifestations. This book is concerned with the intellectual problems associated with those aspects of information transfer most likely to be met by the librarian (using the word in its widest sense); they remain the same no matter what physical means are used.

Current scanning and retrospective searching

Our reader may be mainly interested in keeping up to date with current publications in a subject, in which case our retrieval system must also be up to date. However, because the items referred to are usually easily available, our system need only be a fairly simple guide; if an item looks interesting, the reader can obtain the original without much trouble. On the other hand, the reader may need as much information as can be found regardless of date; in this case, much of the material may be difficult and therefore expensive to obtain, and we need to be much more certain that it will be of use before we attempt to follow up a reader's request. Our information retrieval system must give us enough information about a document for us to be able to decide whether to pursue it or not. Since this second situation is the more demanding, it is the one on which we shall be concentrating in this book, but the more straightforward current scanning should not be forgotten. The contrast between the two approaches is well illustrated by such works as *Current papers in electrical and electronic engineering* and *Electrical and electronic abstracts*, both of which cover the same groups of documents but with two different purposes in mind. There are a number of similar publications covering various subject fields; within the library, current scanning needs are often met by current accessions lists, while the catalogue serves the major function of the retrospective searching tool as far as the library's own stock is concerned.

Selective dissemination of information

In addition to providing facilities for current scanning and retrospective searching, both of which imply that the user takes the initiative, for many years now libraries have themselves taken the initiative by endeavouring to see that readers are kept informed of new materials in their fields of interest. In the public library, this might be on a haphazard, 'old boy', basis, but in the special library it has always been regarded as an important part of the library's function. There are, however, certain difficulties in the way of running such a service successfully, some intellectual, some clerical. The use of a computer can solve many of these problems and enable us to give a more complete and accurate service to our readers.

A system for computer operation was developed by H. P. Luhn of IBM and is still valid today, though it has been modified in some respects. In effect, it involves readers in stating their requirements in the same method of subject description as is used in indexing the library's holdings. If the library uses a thesaurus, then terms will be chosen from this; if a classification scheme, this will be used. Natural language terms may be used, particularly in subject areas such as Science and Technology, where the terminology may be described as 'hard', i.e. well-defined and generally used; the documents may then be regarded as self-indexing, using titles, abstracts or full texts. These reader 'profiles' are fed into the computer together with the similar profiles for new accessions; when the computer finds a match between the two, it prints out a notification, or may use electronic mail facilities to give the reader a more immediate service.

Clerical problems are thus fairly easily solved. The intellectual problems are rather more intractable. A research project showed that perhaps the most pressing difficulty in setting up a viable SDI system was to obtain a valid statement of readers' needs. Users were asked to state their interest profiles, and were sent a selection of articles on the strength of this. At the end of the month they were asked to state which of the articles had been of use, and which article they had read during the month had proved most interesting to them. While the majority of the references notified by the SDI systems were of some value, the 'most interesting articles' were often found to bear little relation to the reader's profile! By asking readers to return the notification form, indicating whether the reference had been of interest or not, a degree of feedback can be obtained which can be used to modify their profiles, but there will never be any means of foretelling the 'wayout' article which may prove of interest.[3]

Despite the difficulties, the IEE developed this work into a satisfactory computer-based system, INSPEC, in which all the operations involved in the SDI service and the production of the various parts of *Science abstracts*, including *Electrical and electronics abstracts*, and *Current papers* . . . are integrated. This is only one of many such services.

While SDI systems may not be able to achieve the impossible, they can function very effectively within a particular organization, and computer processing enables us to extend the benefits to a larger audience. The success of the many services now available has shown that provided the users do their part by stating their needs precisely, a very effective service can be given on a nationwide scale.

In sum, readers will need all the information that we can collect (at least that is the hope of the authors!), but we cannot tell in advance what items of information we are likely to acquire that will be of value to any particular reader. What we have to do is organize our collection in such a way that when we search for information for a reader we do not have to scan the whole contents in order to find what he or she wants, but can go with the minimum of delay to those items which will be of use. To look at it from another angle, our organization must permit us to eliminate what is *not* wanted. This idea introduces three very important concepts: recall, relevance and precision.

Recall, relevance and precision

For any particular reader with a need for information, there will be certain items in our collection which will be relevant. Among these it will be possible to establish some sort of precedence order; some will be definitely relevant, others will be useful, but less so, while others will be only marginally relevant. To take an example, a reader might want information on Siamese cats: in our collections we may have items dealing specifically with Siamese cats, and these will probably be highly relevant. There are however factors other than the subject alone which will influence this; these items may be too detailed, or not detailed enough; they may be written at the wrong level, or in a language which the reader does not understand. The reader's background will inevitably affect any decision as to which items are most relevant. To find more information we may broaden our search: that is, present to the reader those items which, though they do not deal specifically with the subject of the enquiry, include it as part of a a broader subject. In our example, we may find items which deal with cats in general, not just with Siamese cats; or with pets in general, not just with cats. However, we must accept that the more we broaden our search – the more material we *recall* – the less likely it is that any given item will be *relevant.* There is an inverse relationship between *recall* – the number of items we find in conducting a series of searches – and *relevance* – the likelihood of their matching our reader's requirements.

Normally, readers will be satisfied with a few items, so long as these contain the sort of information wanted; that is to say, we need a system which will give us high relevance, even though recall may be low. But there will be situations when readers will require high recall – as much information as possible – even though this means that they will have to look through a lot of items which will turn out to be of little or no value. We need to be able to vary the response of our system to cater for the kind of demand. It is also clear that relevance is a subjective judgement depending on the individual; the same question posed by two different readers may well require two different answers. Indeed, we may carry the argument further. Each document revealed in our search may change a reader's view of what is relevant, so that even a single individual may make varying decisions about relevance at different times.

The problem arises from the fact that readers seek information which they can build into their own corpus of knowledge – their frame of reference – with the minimum of effort, whereas authors present information in a form dictated by *their* frame of reference; each of us has our own frame of reference, so that there will never be an exact match. We have to design our information retrieval systems to optimize the likelihood of being able to match our readers' requests, but accept the fact that they will never be perfect.

The individual view of relevance has led to the concept of *pertinence,* or *utility*. If a document is retrieved in answer to a particular request, its relevance may be assessed by a panel of those skilled in the art,[4] but its pertinence can only be assessed by the originator of the request. In other words, relevance is a consensus judgement, pertinence an individual judgement. Another way of looking at the matter is that a document retrieved in answer to a request may be *useful* to the enquirer, but its utility may change; for example, if we retrieve the same document in a second search, it will have lost its utility the second time round. Its *relevance* will not have changed, but the enquirer's view of it will.

In an experimental situation, such as a study into the effectiveness of different indexing systems, judgements on relevance may be made in advance, for example by examining all the documents in the test collection in relation to all the test questions, as in the second Cranfield project.[5] We can then arrive at an objective view of the success of a system by comparing the results achieved with that system with the predetermined answers. In this situation it is usual to refer to precision rather than relevance. The term precision is used very widely in the literature in preference to relevance, but in this text we shall be using the word relevance when a subjective judgement is involved. The various terms are discussed thoroughly by Lancaster in three articles in the *Encyclopedia of library and information science.*[6] A more recent study shows that the inverse relationship between recall and relevance, first demonstrated in the first Cranfield project,[7] can be mathematically proved. If we want to have improved recall *and* improved relevance, we have to change our search strategy.[8]

We may use Venn diagrams (see Figure 2.1) and set notation to examine these concepts further. If we take as our universe a set of documents L, then in response to any given question there should be a set A of documents which are relevant, where A is a subset of L ($A \subset L$). If we use our information retrieval system to try to find these documents, we shall actually retrieve a rather different set B ($B \subset L$), of which only the subset forming the intersection of A and B ($A \cap B$) will be relevant.
We may now define the two terms recall ratio and precision ratio.

$$\text{Recall ratio} = \frac{(A \cap B)}{A} \qquad \frac{\text{(relevant documents retrieved)}}{\text{(total of relevant documents)}}$$

$$\text{Precision ratio} = \frac{(A \cap B)}{B} \qquad \frac{\text{(relevant documents retrieved)}}{\text{(total of documents retrieved)}}$$

These are usually expressed as percentages by multiplying by 100.

Another term which is sometimes used is fall-out ratio, defined as

$$\text{Fall-out ratio} = \frac{B - A}{A'} \qquad \frac{\text{(retrieved but not relevant)}}{\text{(total not relevant)}}$$

The set $B - A$, consisting of documents retrieved but not relevant, may be regarded as noise, while the set $(A \cup B)'$, documents neither retrieved nor relevant, may be thought of as *dodged*, to use Vickery's term.

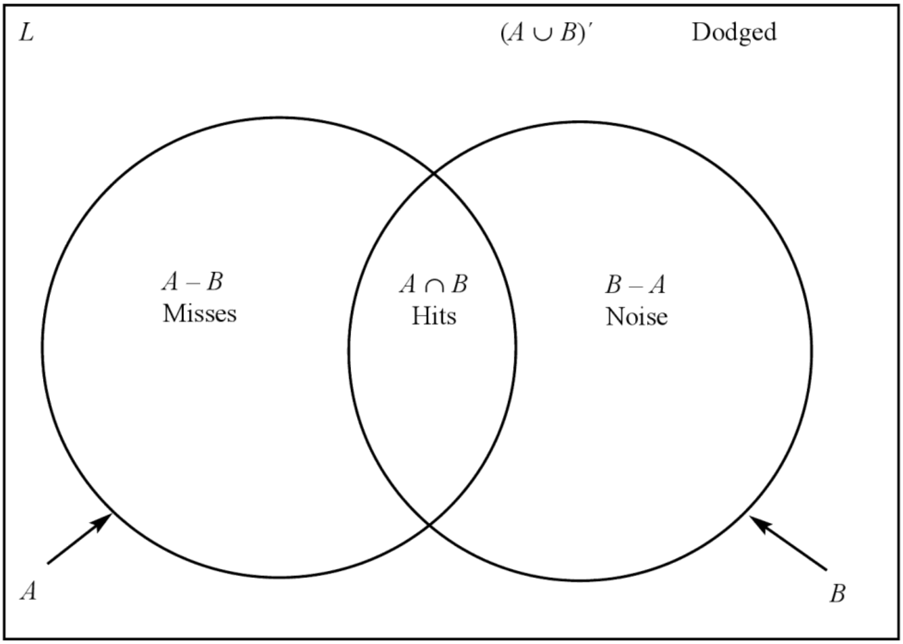


Fig. 2.1 Venn diagram of the retrieval process

Unfortunately, the set *A* is rarely clearly defined; in fact, probably only in the experimental situation can we delineate *A* precisely. In real life there is a grey area, consisting of those documents which *may* be relevant. If we draw a cross-section through *A* and plot this on a graph showing degree of relevance (see Figure 2.2), we get the result denoted APUPA by Ranganathan. U denote the *umbra*, i.e. those documents which are clearly relevant (within the *shadow* of the subject); P denotes the *penumbra* (the 'twilight zone'); and A denotes *alien*, i.e. those documents which are clearly *not* relevant. It is the penumbra which makes it impossible to define *A* clearly, and this means that we cannot use the term precision in this situation. In this text, *relevance* is used to refer to the real-life situation, *precision* to refer to the experimental situation where the set A can be predetermined.

Fuzzy sets

Ranganathan was by education a mathematician, so we should perhaps not be surprised to find that part of mathematical set theory has been used in attempts to quantify his APUPA concept. This is the idea of *fuzzy sets*. In standard set theory, given a universe *L* of which *x* is a member and *A* a subset, we can define a membership function $F_A(x)$ which will be 1 if *x* is a member of *A* and 0 if it is not; in other words, *x* either is, or is not, a member of set *A*. If *A* is a fuzzy set, then $F_A(x)$ may take any value between 0 and 1; in other words *x* is not, ($F_A(x) = 0$); is, ($F_A(x) = 1$); or may be, ($0<F_A(x)<1$); a member of *A*. The concept has been widely adopted since its original publication by Zadeh[9] in 1965, and *Social science citation index* lists sev-

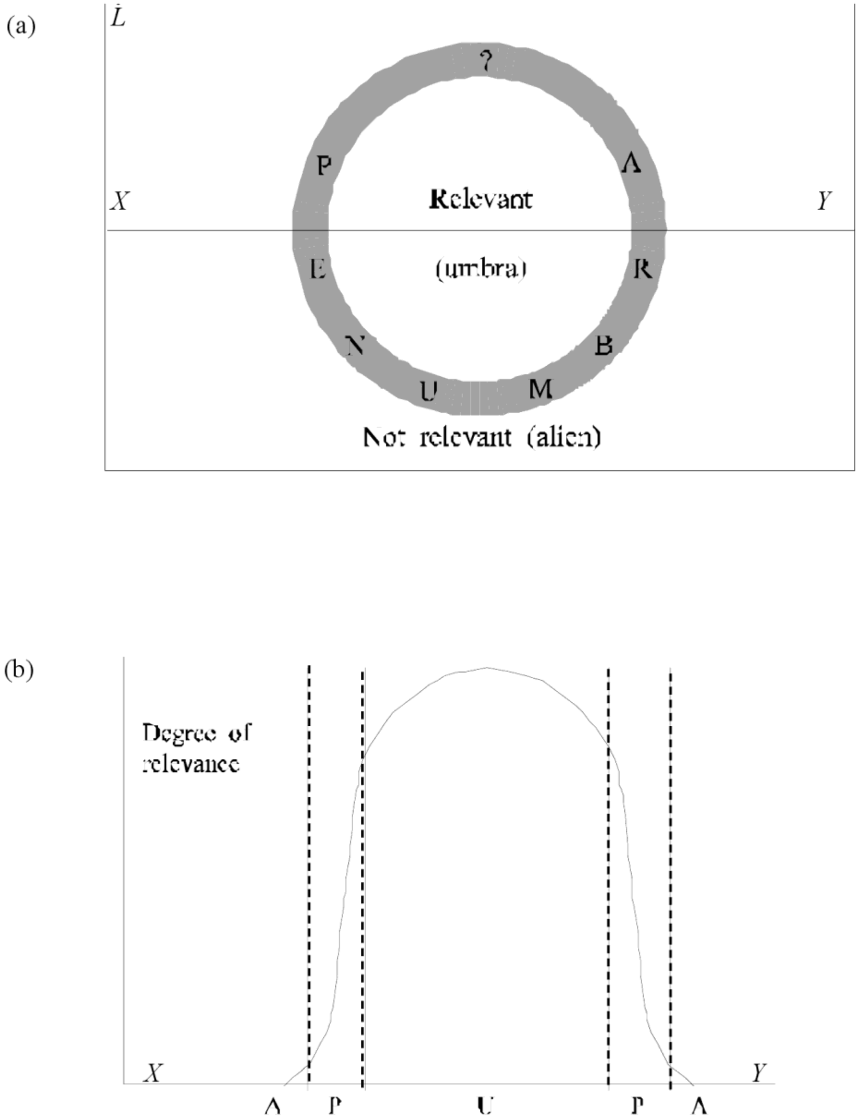


Fig. 2.2 The problem of defining the class *A* of relevant documents:
(a) the boundary between *A* and *A′* is indeterminate;
(b) Ranganathan's APUPA cross-section

eral hundred references citing this, of which only a small proportion are about information retrieval.

The object of applying fuzzy set theory to information retrieval systems is that, by using a computer, we may be able to devise a system that will allow us to rank the documents retrieved in probable order of relevance, instead of the 'either/or' result of a conventional search, which sorts the relevant sheep from the irrelevant goats and ignores the fact that there are a lot of ‘maybe’s’ involved. Systems which rank document output do exist, using a matching function between request and document to rank the output. Other systems use probability functions for the same purpose. It remains to be seen whether fuzzy set theory can produce better results. Some mathematicians are still dubious about the value of the theory as a whole, while other writers have suggested that it is not particularly applicable to information retrieval.

The recall-precision curve

We can display the four classes of document $A \cap B$, $A - B$, $B - A$ and $(A \cup B)'$ in the form of a matrix:

	Retrieved	Not retrieved	Total
Relevant	$A \cap B$	$A - B$	A
Not relevant	$B - A$	$(A \cup B)'$	A'
	B	B'	L

From this matrix, we may see why some writers prefer to compare relevance ratio and fall-out ratio rather than relevance ratio and recall ratio. The latter two have the same numerator ($A \cap B$), whereas fall-out ratio uses the complementary section of the matrix, giving a different numerator and denominator from either of the others (($B - A$) and A' instead of ($A \cap B$) and A or B).[10] However, recall and relevance are more usual. We can take the results of a number of tests and use these to plot a graph of recall ratio against precision ratio. Ideally, of course, our graph would be concentrated into the 100% recall and 100% precision corner, but in practice we obtain a curve of the kind shown in Figure 2.3.

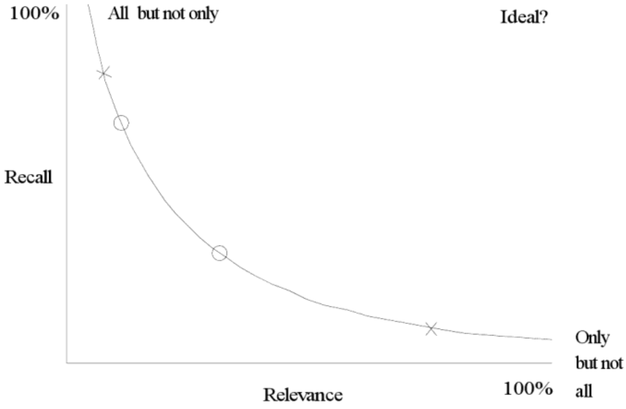


Fig. 2.3 The recall–relevance/precision curve
Flexibility of system: ability to select the appropriate operating point on the recall-relevance curve
System X–X is more flexible than system O–O (X and O mark limits of operating parameters)

The implication of this curve is that if we try to improve recall, we can only do so at the expense of precision, and conversely that if we try to improve precision, we can only do so at the expense of recall. We owe the terms recall, relevance and precision to the Cranfield Project, but the same idea was expressed some years earlier by Fairthorne in the phrases 'All but not only' (ABNO) and 'Only but not all' (OBNA).[11] One measure of the effectiveness of an information retrieval system is the freedom with which one may move from one part of the recall-precision curve to another; for example, if our first search does not reveal all the information we want, can we increase recall by moving up the curve (and thereby sacrificing a degree of precision)? If our first search reveals an overwhelming amount of information, can we increase precision and thus reformulate our search strategy to give lower recall? Ways of altering our search strategies to give such changes will be discussed in due course, when it will become apparent that not all systems have the same degree of flexibility in this respect.

We should be cautious about accepting the recall-precision curve unquestioningly. As Cleverdon himself has pointed out, it represents the average performance of any given system, and this may vary quite considerably in particular situations. For example, in the MEDLARS evaluation study Lancaster found that the system was operating on average at about 58% recall and 50% precision, retrieving an average of 175 documents per search. To achieve 85% to 95% recall would have meant retrieving an average of some 500 to 600 documents per search, with a precision ratio of about 20%. However, if we examine the results of individual searches, we find that in some cases 100% recall was achieved with 100% precision, while in others both recall and relevance were zero! Furthermore, while it may not be possible to alter the response of a system to a particular request – if we try to improve recall or precision we can only do so at the expense of the other – in practice, once we start to obtain documents from the system in response to a request, the feedback we obtain may well enable us to modify our request in such a way as to improve both recall and precision.

From the point of view of the user, it is usually relevance rather than recall which is desirable. The majority of enquiries can be satisfied by providing no more than half a dozen documents, providing they are all useful, and it is only in the minority of cases that a high recall figure is necessary. For example, the Patent Examiner needs to be sure that any relevant documents that may exist are found, since prior publication is grounds for invalidating a patent; *one* such document is enough, but the system must be such that the Examiner is certain to find that document if it exists. On the other hand, the casual enquirer would not normally need the same degree of certainty, and would rarely wish to be presented with an enormous pile of documents in response to a request.

There is also the point that many documents are repetitive: they do not add anything significant to our knowledge. R. Shaw examined a collection of documents on milkweed in detail, and came to the conclusion that all the information to be found was contained in 96 of the total of 4000![12] Of course, in order to identify the 96 he did have to examine the 4000, but this does reinforce two points. The first of these is that even literature which is claimed to be original may duplicate previous work;

the second is the importance of the intermediary whose task is to act as a filter between authors and users. There is scope for a great deal more analysis of the content of documents, as opposed to their unquestioned supply to the user who then has to sort the wheat from the chaff. It has been suggested that we are seeing a publication explosion rather than an information explosion; perhaps librarians could solve some of the problems of recall and relevance by encouraging some kind of literary contraception.

We should however not overlook the fact that there is a genuine need for a certain amount of duplication. New ideas need to be disseminated at more than one level; what would be ideal for the author's peers would probably leave the lay person no wiser, and an interpretation at a simpler level is required. With the modern trend towards increasing specialization, we may even have various levels of 'lay person', requiring a series of levels of interpretation ranging from the original article in a learned journal to the TV presentation and newspaper report. There is also the problem of publication in languages other than the original. Depending on the purpose of our information retrieval system, we may need to store information at more than one level, and ensure that we have the means of selecting the right level for any particular user.

Blair has put forward an interesting and useful concept to clarify the 'normal' attitude to recall and relevance, that of the *futility point.*[13] A relevance ratio which might be quite acceptable in a small system may be totally unacceptable in a computer-based system with access to millions of references. Two terms have been coined to characterize a user's reaction to the result of a search. To satisfy the user, a search should retrieve a small enough set of documents for the searcher to be willing to browse through them to find those which are relevant. The number of documents through which searchers are willing to browse before giving up in disgust is their futility point FP. If a searcher's FP is *n*, then they will be willing to look through n documents of the set B in Figure 2.1 before giving up without finding the document(s) wanted from set A. However, there is also an 'anticipated futility point' *m*, which is the size of the set that a searcher is willing to *begin* browsing through; for example, it is obvious that in the vast majority of cases the announcement that a search has retrieved 1,200 references will not rouse any great enthusiasm in the searcher. Based on observation, Blair suggests that in practice *m* is of the order of 30; with many databases now containing millions of references, we may have to rethink our ideas on acceptable relevance levels.

Confirmation of the concept of the futility point comes from a study by Lantz on the London University Computer Information Service LUCIS.[14] Users were asked to respond to questionnaires on over 2000 requests; the first asked them to estimate the number of relevant references retrieved, the second asked how many they had actually read. The responses showed that the more relevant documents were retrieved, the smaller the proportion of them that were actually read. When the number relevant was 10, the number read was about 5; when the number relevant was 100, the number read was about 28. The proportion varied according to the subject field; for engineering, the number read reached a limit at about 10, for medicine at about 60. For the social sciences, the field observed by Blair, the figure was about

30. Mathematical analysis showed that the curve of number read y against number relevant x could best fit an exponential expression:

$$y = a(1 - e^{-\mathbf{bx}})$$

where the mean value for a is 34.1 and for b is 0.017. As $x \rightarrow \infty$, $e^{-bx} \rightarrow 0$, so that $a(1-e^{-bx}) \rightarrow a$: the mean maximum number read is then about 34. These are average figures (two readers in fact reported that they had read over 200 references) but the similarity to Blair's figure is too striking to be ignored. The University of London Library has of course very good resources for users to consult; in libraries with fewer resources, it is probable that the proportion read would be influenced by the availability factor. Users normally prefer materials that are immediately available, even if they are not the 'best possible', to those which have to be obtained from elsewhere, with an associated delay. In a library with fewer resources, it is likely that the proportion of relevant items actually read would be lower than was found in this study.

Despite the need to be aware of these findings, it is clear that recall and relevance (precision) are valuable concepts in the study of any information retrieval system. They are, however, not the only criteria by which a system may be judged. We can now consider some of the factors affecting recall and relevance, and some of the other important aspects of IR systems.

Probability of error

Indexers are human; so are users, and also the keyboard operators who produce computer-readable text. All are thus liable to make mistakes. Our system should be one which reduces the probability of error as far as possible. Research by telephone engineers some years ago[15] showed that the probability of incorrect dialling began to rise steeply as the length of the number increased to nine or more digits. (A look at trends in telephone numbers can be depressing.) If a system uses numbers for coding, for example a classification scheme, mistakes become more common if the length of the notation grows beyond the limit indicated above. Even if we use words, the likelihood of error still exists; for example, many English users of MEDLINE must have been infuriated by the inability of the system to spell even simple words such as Haemoglobin and Labour correctly. This problem is compounded in Australia, where there is an ambivalent attitude towards words ending in -our; the preferred spelling as shown in the *Macquarie dictionary* is -our, but one major political party is the Labor Party, and over a century ago one port in South Australia was named Victor Harbor!

Errors will have an effect on relevance, in that we shall get answers which are wrong; they also affect recall, in that we shall miss items that we ought to find. We should therefore try to ensure that the system we use does not have a built-in tendency to increase human error. Just as faults in a telephone line introduce audible noise and thus interfere with our reception of a message, so errors in an IR system introduce their own particular kind of noise. The fewer the errors, the less will be the noise from this source.

Ease of use

Another source of noise is the ease with which an IR system may be used. Whatever system we choose to use, there are two groups of people who must find it helpful: those responsible for the input, i.e. the *indexers*, and those trying to obtain an output, i.e. the *users*. At the input stage, how much skill do indexers need to be able to use the system? Does it help to overcome deficiencies in their understanding of the subjects dealt with? Despite their best endeavours, indexers are not omniscient! Similarly, users often find it difficult to express their needs exactly; does the system help them to formulate a satisfactory search despite this? Is the physical form of output acceptable? A system which presents the user with a set of documents, or at least abstracts, is likely to be more popular than one which gives merely a string of numbers.

Specificity and exhaustivity

There are several other factors which affect the overall performance of an information retrieval system and its potential in terms of recall and relevance. First, we may consider *specificity*: the extent to which the system permits us to be precise when specifying the subject of a document we are processing. The higher the specificity, the more likely we are to be able to achieve high relevance, and conversely, with a system that permits us only limited specificity we are likely to achieve reasonably high recall but correspondingly low relevance. In the example quoted earlier, if our system did not permit us to specify *Siamese* cats, we should have to look through all the items about cats before we could find out whether we had anything on that particular breed. Further, if the second item we found did relate to Siamese cats, there would be no guarantee that this would be the only one, or that any others would be found alongside it. If specificity is lacking, we are in fact reduced to the kind of sequential scanning that is necessary if our collections are not organized at all – though of course we have reduced the amount of material that we have to scan by partially specifying its subject content. If we are to obtain the maximum amount of control over our searching, the system must permit us to be precise in our specification of subjects; in fact, our specification should in every case be coextensive with the subject of the document. If we need to increase recall, we can always ignore part of our specification, but we cannot increase relevance by adding to it *at the search stage*. It is very important to keep clear the distinction between the *input* to our system (i.e. the specifications of the documents we are adding), and the *output* (i.e. the results of the searches we perform among these specifications). We cannot add to the input at the output stage; anything omitted at the input stage will remain outside the system, and will have to be replaced by sequential scanning of an unnecessarily large output. (We may however be able to use a systematic approach to help us formulate our search strategy, even though it is not part of the system used to specify the input, while the use of computers may enable us to get round the problem by using full texts as input and thus transfer all our ‘indexing’ operations to the output stage. These ideas are discussed in more detail later.)

Specificity is a function of the system, but another important factor, *exhaustivity*, is the result of a management decision. This is the extent to which we analyse any given document to establish exactly what subject content we have to specify. We may distinguish between the overall theme of a document, and the subthemes which it may contain; for example, a description of a scientific experiment may be concerned overall with the purpose and results, but it will probably also contain a description of the apparatus used. In a large general library we may content ourselves with specifying the overall themes, giving perhaps an average of one to one and a half specifications per document, whereas in a small special library, anxious to exploit the stock to the maximum advantage, we may wish to index subthemes as well, giving perhaps dozens of specifications for every document. This is known as *depth indexing*, as opposed to the *summarization* of the first method. Depth indexing might indicate that a book is about Dryden, Wycherley, Congreve, Vanbrugh and Farquhar, while summarization might say that it was about Restoration drama. It would clearly be very difficult if not impossible to index *all* the subthemes in a document in a library catalogue or a bibliography (the index to the present work contains nearly 1000 entries), so depth indexing is usually carried out in libraries where the needs of the readers can be foreseen fairly clearly; often depth indexing is applied to technical reports and similar documents which are relatively short and are therefore manageable. Some work has been done on the feasibility of using book contents pages and indexes to increase greatly the exhaustivity of indexing possible in online systems with a minimum of effort.[16] As a high proportion of the books produced today are computer-typeset, machine-readable versions are easily obtained.

Depth indexing involves the indexer in the exercise of judgement as to which themes and subthemes are worth noting. In choosing between depth indexing and summarization, the decision is ours and not a function of the system. However, there is a link between exhaustivity and specificity in that there is no point in increasing exhaustivity unless the system being used has adequate specificity.

A moment’s thought should show that whereas specificity is a device to increase relevance at the cost of recall, exhaustivity works in the opposite direction, by increasing recall, but at the expense of relevance. A device which we may use to counteract this effect to some extent is *weighting*. In this, we try to show the significance of any particular specification by giving it a weight on a pre-established scale. For example, if we had a book on pets which dealt largely with dogs, we might give PETS a weight of 10/10, and DOGS, a weight of 8/10 or less. If it gave some information about dogs, but not much, we might give DOGS a low weighting of 2/10. The reader wanting high relevance now knows that this particular item can be ignored, at least for the time being, while the reader wanting high recall will have no difficulty in finding it. Search terms may be weighted in the same way, and the SMART system incorporates means for using feedback from the user’s relevance judgments to amend the search strategy by altering the weighting given to the various search terms.

Time

Indexing takes time; searching takes time. By increasing our effort at the indexing stage – the input – we may well be able to reduce the amount of time we have to spend at the output stage in searching. On the other hand, in any given library situation, a proportion (which may be high) of the documents indexed will never be sought, and the effort used to index them will be wasted; if we concentrate our effort at the output stage by keeping our indexing to a minimum (for example by using abstracts or even whole texts for the input instead of subject specifications) and then perform complex searches to find relevant items, we can argue that we are eliminating a large amount of unnecessary work. As has already been mentioned, users cannot always specify exactly what it is they want, so any search will be a dialogue between the user and the system; the results of a first search will be used to modify or refine the question so that further searches can be performed until such time as a satisfactory end point is reached. By concentrating our efforts on the searching rather than the indexing, we do not hamper this dialogue in any way, but we can very easily make use of such feedback in planning future search strategies. In a system where the effort is concentrated on the input, this may not be quite so easy, as the output cannot affect the input retrospectively; to take account of experience gained in searching we may have to re-index some items – i.e. increase the input effort still further.

At present, nearly all systems involve large amounts of input effort rather than transferring this to the output stage. The complicated search strategies necessary with the latter technique, together with the very large amounts of storage necessary for the input, have only recently become available as computer technology has advanced, and there remain many situations where this approach is still either impractical or not cost-effective. Some of the SDI services referred to earlier rely on abstracts or even titles on their own, and the consequences of this in terms of search strategy and profile construction are discussed in more detail later. Searching of full texts is now becoming commonplace, but is likely to present problems of relevance, since exhaustivity is obviously 100%.

Iterative and heuristic searching

The idea of a dialogue between users and system is worthy of further examination. As has been pointed out above, users often find it difficult to express their needs precisely. In a conventional library, searches may be carried out by users, or by the librarian acting on their behalf. When they are carried out by the users themselves, searches are usually modified as they progress; each relevant document found tends to influence the user's decision as to what further information is required. In many cases the clarification that results as searches are pursued leads to a situation where users finish up with objectives rather different from those that they started out with. Such a search, where the course of events is modified continuously in the light of knowledge being gained, may be described as *heuristic*. If on the other hand the search is carried out by the librarian, this continuous modification is not possible, since modifications of the librarian's knowledge do not affect the user. For this reason it is usual for the librarian to perform a first search and present the results to the user; the search strategy may then be modified in the light of the proportion of relevant documents among those resulting from this search. A second search may then be performed, and the process repeated until the user has what is needed. This kind of search, which is modified not continuously but at intervals, may be described as *iterative*. Both heuristic and iterative searches require interaction between user and results, but heuristic searching eliminates the time delay between receiving the result of a search operation and using it to modify the search procedure.

Many information retrieval systems do not permit heuristic searching, whereas the conventional library card catalogue does. The importance of this should not be overestimated, but it is obviously a point to be considered when we are trying to estimate the relative value of different systems.[17]

Browsing

We have assumed so far that the purpose of our system is to make it possible to find information on demand – that the users will approach it with some definite objective in mind, even though they may not, to begin with, have clarified this. However, this is by no means always the case; there will be many occasions when readers will approach the collection without any particular need in mind but wishing instead to be able to select items at random. To help in this situation, our system should permit *browsing*; a reader should be able to follow a casual train of thought as well as a planned search. As was pointed out in the discussion of SDI, it is often an item which does *not* fit our existing patterns of interest which proves to be the most interesting; many of the most significant scientific discoveries have arisen as the result of *serendipity* – 'the faculty of making happy and unexpected discoveries by accident' – and a system which excludes this possibility might prove to be *too* successful in matching readers' expressed needs!

Cost

Many of the factors affecting information retrieval systems are cost factors. We have to balance the cost of so organizing our libraries that we can find information when it is required, against the cost of not finding it at all, or finding it too late for it to be of use. In libraries serving industrial firms, for example, the cost of not finding information may be high; this is why 'hard headed businessmen' add to their overheads by paying for extensive library services. (The term 'library services' is here taken to include those denoted by the more elite term 'information services'). On the other hand, public libraries have in the past tended to regard the exploitation of the information in their stocks as very much less important than its provision, because the cost to the community at large if one of its individual members fails to find information is considerably less than the cost of organizing the material adequately. However, it is now being realized that the cost to the community of wasted information is in fact very high in terms of international competition, and more effort is being devoted to providing adequate services. We still have to find out a great deal about the cost effectiveness of various methods of organizing information, though we are beginning to learn something about their comparative efficiencies as systems. Despite our relative ignorance we must not ignore cost factors altogether, but they can usually only be studied in detail in a particular set of circumstances, and will therefore only be indicated in general terms in this text.

Modern trends in the evaluation of cost significance have been towards the idea of cost effectiveness. Most of the sophisticated devices developed in recent years have been aimed at improving relevance: reducing the number of unwanted documents revealed by a search, and thus reducing also the time taken to scan through the results and select those which are of use to us. However, if it costs more to use a sophisticated system for indexing than it would cost to look through the output of an unsophisticated system, there is no point in using the more advanced system. We have also to bear in mind that a relevance level which might be tolerable in a small system might well be quite unacceptable in a large nationwide mechanized system. If a search reveals ten documents, four of which are useful, this is not too bad; but if we have the same level with a collection a hundred times as large, we might well boggle at the thought of discarding 600 documents from a total of 1000. As yet only a limited amount of research has been carried out into this aspect of information retrieval, but it is obviously a field that is likely to be explored in more depth in the future, particularly with the development of mechanized systems.

Problems of linear order

Knowledge is multi-dimensional: that is to say, subjects are related one to another in many different ways. In the example quoted earlier, it was assumed that Siamese cats were to be considered as pets, but it is obvious that they can be regarded in many other ways – as a branch of the zoological class *Felidae*, or as originating in a particular part of the world, to name but two approaches. However, when we try to arrange items in our library or catalogue, we find that we are restricted to a linear, unidimensional, sequence, just as we are if we are reading a book. We cannot *display* multiple relationships and must therefore find some other means of showing them. If we have a book with no contents list and no index, the only way we can find a given item in it is to read it through. We have only one means of access; sequential scanning. However, we can overcome this problem by providing multiple access through the contents list and index, which permit us to go direct to the information we require; but the text of the book continues to display its information unidimensionally. The sequence in the book is chosen for us by the author and we cannot alter it, though we may to a large extent minimize the effect by adequate signposting in the form of indexes and guiding.

We face exactly the same problem in organizing the information in our libraries. We can provide a sequence which we hope will be helpful to our readers, just as an author does, but we must recognize the need to cater for other modes of access. We must also realize that without these secondary modes of access we can only find information in one way, unless we are prepared to revert to sequential scanning. A simple example will demonstrate this in relation to a familiar tool, the telephone directory. These directories are arranged according to the surnames of the subscribers, set out in alphabetical order. Provided we know the subscriber's name, we can find the telephone number without much trouble, but we cannot perform the operation in reverse; we cannot find out the name of a subscriber whose number we know, unless we are prepared to look through the directory until we find it. To overcome this we can have a second sequence, arranged this time by number; but we still cannot find the number of a friend if we only know his or her forename and address.

The problem is of course largely an economic one. We do not set up multiple sequences of books and other items in our libraries because it would cost too much to try to arrange a copy of a book at every point in the library where it might be related to other items. Nor can we afford to make multiple sequences in bibliographical tools which have to be printed and distributed. We might perhaps make several sequences in our records within the library, but even this will prove very expensive if we are to be consistent and comprehensive. However, just as we can overcome the problem in a book by providing multiple access through subsidiary sequences which lead us to the required points in our main sequence, so we can do the same thing in our information retrieval system. Different systems will permit us differing degrees of multiple access; the more flexible a system is in this respect, the more likely it is to be of value. This is perhaps the major advantage of computer-based systems. Once the information is stored in the machine, it is possible for us to manipulate it in virtually as many ways as we wish; we have all the flexibility which has long been recognized as desirable but unattainable with manual systems.

Literary warrant

No matter what our system may be, the information in it must be a function of the input; that is to say, our systems must take account of the relationships between subjects shown in the items we are indexing. We may in addition built into it relationships between subjects of which we are aware *a priori*, through a study of knowledge *per se*, but if we restrict ourselves to a study of knowledge alone without taking into account knowledge as it is presented in recorded form, i.e. information, we shall find ourselves unable to specify subjects precisely. In other words, we are concerned with the organization of information rather than the organization of knowledge on its own. The term *literary warrant* is used here to denote that our system must be based on the information we put into it rather than on purely theoretical considerations. (As this term is widely used in this context, it is retained, even though computer-based systems may well hold information which is not available in any other form, and conventional systems may well hold information relating to audiovisual materials which do not fit the usual definition of 'literary'.)

There is another aspect to this particular question. It is the output of the system which is important, since this is the whole purpose of the system. But we cannot know in advance what output will be required, at least not with any degree of precision, though we may be able to form an intelligent guess on the basis of past experience. So although it would be desirable to build up our system in such a way that it matched the required output, we are unable to do this since we do not know what the required output will be. We are obliged to use the input as our basis for building up the system, adding to this whatever is suggested by studies of knowledge outside the system. If we restrict ourselves to studies of knowledge outside the system we shall, by ignoring the input, be removing our system one stage further from the required output. In any subject area there will be an accepted corpus of knowledge, but each document we index may modify this; literary warrant implies a system that is able to accept this kind of change.

There is perhaps a danger that we may take a negative attitude to literary warrant: exclude from our system the possibility of catering for subjects which have not as yet appeared in our collections. This danger is usually associated with the older kind of enumerative system described below, but there have been more recent examples to demonstrate the problems that arise if we deliberately make our system a static one. Hospitality to new concepts as they are revealed by our collections is vital if we are to maintain the desired level of specificity.

The term literary warrant was used by Wyndham Hulme to denote a rather different kind of idea, though one basically similar.[18] He considered that if we have a document entitled, say, *Heat, light and sound*, then that represents a subject for which we should make provision in our system. However, most of these are not genuine subjects but aggregates of subjects resulting from the bibliographical accident of being bound within the same pair of covers. This situation should not be confused with that of a genuine interaction between subjects, e.g. the effect of heat on sound (discussed in more detail on pp.107–110); this is a different kind of situation for which we do have to make provision. Hulme's use of the term is rarely found now, though his ideas were largely reflected in the practice of the Library of Congress, and have indeed developed into the modern theory already outlined.

Heading and description

We use the terms in our indexing system to name the subjects of the documents in our collection, but obviously a user who has found the correct subject description will require in addition some details of the documents to which that description applies. We can therefore divide an entry in the system into two parts, the heading and the description.

The *heading* is the subject description which determines whereabouts in the sequence we shall find any given entry. (The present work is restricted to considerations of the subject approach; in a full catalogue, headings will include names of authors and titles, as well as subjects.) In an alphabetical system, headings will consist of words, while in a systematic arrangement it is the notation – the *code* vocabulary – that is used for the headings. We need to distinguish between two kinds of heading. We have those which are 'preferred' terms, in the sense that we use them to lead us directly to information – a catalogue entry, or a book on the shelf. These form the *index vocabulary*. We also have those which are 'non-preferred', for examples synonyms we have decided not to use, except to lead us to terms in the index vocabulary. These non-preferred terms, together with the index vocabulary, form the 'entry vocabulary'.

The *description* is the part of an entry which gives us information about a document, and will therefore contain all those factors which serve to *identify*. There are various sets of rules for the compilation of document descriptions, e.g. those in the

Anglo-American Cataloguing Rules, or the International Standard Bibliographical Description, but for our purposes here we need only note their existence. The presence of a document description enables us to make a useful distinction: a *subject entry* consists of a heading from the index vocabulary together with a document description, while an *index entry* or *cross-reference* leads us from a heading with no document description to an entry. The heading from which we make a cross-reference may be one which appears only in the entry vocabulary, in which case the reference is one leading us from a heading not used to one which is used; or it may appear in both entry and index vocabulary, in which case the reference is one linking two headings which are both used, to show some kind of relationship.

It should be noted that in some systems the descriptions may be in the form of a number (an accessions number or document number) rather than the detailed information about author, title, imprint and so on which we find in, for example, a library catalogue. The links between related headings may form an integral part of the main sequence, as in a dictionary catalogue; part of a subsidiary sequence, as in as classified catalogue; or quite separate, as is usual in post-coordinate systems. These points will be clarified in due course; at this stage it is important to realize that these features, like the others in this chapter, are common to all information systems. Their presence or absence can make a great deal of difference to the ease with which we can retrieve information.

Term entry and item entry

The preceding section implies that we make entries for a document (which we identify by its description) under each of the appropriate headings, and file these in the correct place in our alphabetical or classified sequence. A system which works in this way is called a *term entry* system, and a card catalogue using unit cards is of this kind. It is possible to take the opposite approach, and make a single entry for each item, using a physical form which permits access to the entry from all appropriate headings. Such a system is known as *item entry*, and is used in computer-based systems. It implies a main sequence of entries supported by one or more indexes, and is discussed more fully in Chapter 5.

Separation of intellectual and clerical effort

In any system, part of the work involved will be intellectual and part will be clerical. Deciding what the access points to a particular document should be is an intellectual operation in assigned indexing systems, discussed in later chapters, but the actual mechanics of placing an entry in a file is not. Similarly, in searching we need to make an intellectual decision as to which words or headings are likely to give a satisfactory answer to an enquiry, but the task of displaying the results of the search does not involve intellectual effort.

The distinction has become important with the use of computers. Computers can perform clerical tasks very well, they are much faster and more accurate than human beings, provided they are given the right instructions. At present, we do not know enough about the way in which the human mind works to be able to give computers the right instructions to enable them to perform intellectual operations; these must still be done by human effort. However, studies such as that carried out by E. J. Coates in the computerization of the production of the then BTI showed that many of the operations which had been thought of as intellectual could be reduced to an algorithm suitable for computer operation. It is clearly advantageous to transfer as much routine work to the computer as we can, to enable us to concentrate on the intellectual tasks; by doing so we can only improve our service to users.

References

1 Vickery, B. C., *Techniques of information retrieval*, London, Butterworths, 1970. Chapters 1 and 2.

2 Orwell, G., *1984*, London, Secker & Warburg, 1949.

3 Kemp, A., *Current awareness services*, London, Bingley, 1979.

4 The phrase 'one skilled in the art' is commonly used in patent specifications to denote someone who has a sufficient working knowledge of the existing procedures to be able to utilize the invention being patented.

5 Cleverdon, C. W., Mills, J., and Keen, E. M., *Factors determining the performance of indexing systems*, Cranfield, Aslib–Cranfield Research Project, 1966, 2v in 3.

6 Lancaster, F. W., 'Evaluation and testing of information retrieval systems', in *Encyclopedia of library and information science*, **8**, 1972, 234–59.

Lancaster, F. W., 'Pertinence and relevance', in *Encyclopedia of library and information science*, **22**, 1977, 70–86.

Lancaster, F. W., 'Precision and recall', in *Encyclopedia of library and information science*, **23**, 170–80, 1978.

Swanson, D. R., 'Subjective versus objective relevance in bibliographic retrieval systems', *Library quarterly,* **56** (4), 1986, 389–98.

For a recent detailed review of the question of relevance see Schamber, L., 'Relevance and information behavior', *Annual review of information science and technology,* **29**, 1994, 3–48.

7 Cleverdon, C. W., *Aslib Cranfield Research Project: report on the testing and analysis of an investigation into the comparative efficiency of indexing systems,* Cranfield, College of Aeronautics, 1962.

8 Buckland, M. and Gey, F., 'The relationship between recall and precision', *Journal of the American society for information science*, **45**, 1994, 12–19.

9 Zadeh, L. A., 'Fuzzy sets', *Information and control,* **8**, 1965, 338–53.

Robertson, T. E., 'On the nature of fuzz: a diatribe', *Journal of the American Society for Information Science*, **29** (6), 1978, 304–7.

Cerny, B., 'A reply to Robertson's diatribe on the nature of fuzz', *Journal of the American Society for Information Science*, **30** (6), 1979, 357–8.

Bookstein, A., 'Probability and fuzzy set applications to information retrieval', *Annual review of information science and technology,* **20**, 1985, 117–51.

10 I am indebted to E. M. Keen for drawing my attention to this point in ref. 5 above.

11 Fairthorne, R., ‘Automatic retrieval of recorded information’, *Computer journal*, 1958, 36–41.

12 Shaw, R., *private communication*, quoted by Cleverdon in *Journal of documentation*, **30** (2), June 1974, 174.

13 Blair, D. C., ‘Searching biases in large interactive document retrieval systems’, *Journal of the American Society for Information Science*, **17** (3), 1980, 271–7.

14 Lantz, B. E., ‘The relationship between documents read and relevant references retrieved as effectiveness measures’, *Journal of documentation*, **37** (3), 1981, 134–45.

15 Conrad, R. and Hille, B. A., ‘Memory for long telephone numbers’, *Post Office telecommunications journal*, **10**, 1957, 37–9.

16 Atherton, P., *Books are for use: final report of the Subject Access Project to the Council on Library Resources*, Syracuse, NY, Syracuse University, 1978.
Cochrane, P. A., *Redesign of catalogs and indexes for improved online subject access: selected papers of Pauline A. Cochrane*, Phoenix, AZ, Oryx Press, 1985.

17 Lancaster, F. W., ‘Interaction between requesters and a large mechanized retrieval system’, *Information storage and retrieval*, **4** (2), 1968, 239–52.

18 Hulme, E. Wyndham, *Principles of book classification*, London, Association of Assistant Librarians, 1950 (AAL Reprints No 1). Originally published in the *Library Association record*, 1911–1912. Included in *Theory of subject analysis . . .*

Chapter 3

Derived indexing 1: Printed indexes

As we saw in Chapter 1, we have to encode the subject of a document in order to place the document itself or our records of it in our store. This means that we must in some way be able to specify the subject. How can we establish the subject of a document so that we can specify it? We do not usually have time to read all the documents we add to stock, and in any case we might not understand them if we did. We may use short cuts: the contents page, preface or introduction, or publisher's blurb on the cover for a book; or an abstract if we are looking at a journal article or technical report; or the claims for a patent specification. All of these will give some indication of the subject and will suggest certain lines of thought if we want to pursue the matter further, for example in a dictionary or encyclopedia.

We may decide that in the interests of economy we will rely solely on information which is *manifest* in the document, without attempting to add to this from our own knowledge or other sources. This is *derived indexing*, that is, indexing derived directly from the document. We can begin by studying some of the ways in which derived indexing has been used to produce printed indexes, particularly in computer-based systems. These are now often found in online systems, but the principles remain the same.

We have seen that it is possible to distinguish between intellectual and clerical effort involved in an IR system, and computers enable us to carry out the clerical operations at high speed. Derived indexing reduces intellectual effort to a minimum, and is thus well suited to computer operations, which can enable us to get a variety of outputs from the one input. We may find that we are able to produce some forms of output that can be produced manually but are usually not attempted because of time and cost factors.[1]

Title-based indexing

There is of course one part of a document in which authors themselves usually try to define the subject: the title. In many cases this will give a clear indication of what the document is about, though we find cases where the title still leaves us in some doubt, and others where the title is obviously meant to attract attention rather than inform us about the subject. In the first category we might place *The development of national library and information services*, or *Early Victorian New Zealand*; in the second, *The design of steel structures* (Buildings? Bridges? Or all of them?), or

Steps in time (Fred Astaire's autobiography!); and in the third *Men in dark times* (a collection of biographies of men who died in the twentieth century, including Bertolt Brecht), and 'Waterfalls and tall buildings', which turns out to be a review of the *Guinness book of records*! Authors tend to generalize in the titles they select, and though they usually try to find titles which are unique to their own work, this is not always the case. There are for example several books with the rather vague title *Materials and structures*. The title *Malice in Wonderland* used by Nicholas Blake for a detective story was also used for a film starring Elizabeth Taylor. So in general we find that searching for specific titles will give low recall, though probably high relevance at the same time, but occasionally will produce false drops – titles which match the specification but are not in any way relevant.

If we consider titles given to 'serious' works, we will often find that works on the same subject have titles containing the same significant words – *keywords* – which can be used as a basis for information retrieval, e.g.:

Manual of library classification
Library classification on the march
Introduction to library classification
A modern outline of library classification
Prolegomena to library classification.

The use of keywords to give various kinds of index is well established, but has been emphasized in recent years by the use of computers to manipulate the terms.

Catchword indexing

Catchword indexing has been used for many years in bibliographical tools, particularly those produced by publishers, where it has provided a cheap and reasonably effective means of subject access to the titles listed. The titles are manipulated to bring the significant words to the front, giving perhaps two or three entries per title. The editor selects the words to be used, and the entries are generated manually. The technique was also used to produce the indexes to periodicals such as *Nature*. With the computerization of these kinds of index, catchword indexing has in effect been dropped in favour of other formats, but it will be found in many older reference tools.

Keyword in Context (KWIC) indexing

One particular form of catchword indexing was adapted for computer production by H. P. Luhn of IBM.[2] Each significant word becomes an entry point, but instead of appearing at the left-hand side of the page, the keyword appears in the middle, with the rest of the title on either side. One important point arises immediately. In catchword indexing, an editor selects the significant terms, but a computer cannot recognize the significance of a given word; instead, we have to construct a list of words which are *not* of value for indexing purposes: a *stop list*. The computer is then programmed to delete any entries which might arise under these terms. It is usual to have a fairly short list of terms which are obviously of no value as index entries –

the *articles* a, an, the; *prepositions* on, of, in; *conjunctions* and, or; *pronouns* he, she, my; and so on – and then add a rather larger list, based on experience, of words which are unlikely to be of any value. The New Grolier multimedia encyclopedia has a stop list of 132 words, including bibliography, which might not please librarians! Other terms considered to be unsought are words such as although, begun, can (which might be unhelpful to a can-maker!), different, etc. The list may be edited from time to time to take account of 'fashions' in terminology and of changes in the subject coverage of the collection. This will still leave us with words which occur too rarely for us to include them in a stop list; when they do occur, they will give rise to entries in the index, but these will be few enough not to bother the user. It is better to have a few entries that are not useful than to omit some which would be.

KWIC techniques were at one time popular as a cheap and quick means of producing indexes (KWIC is an unusually apposite homophonic acronym), but with more sophisticated methods now available, they have been largely superseded. One place where such an index is still often to be found is in a thesaurus, where the *rotated index* is used to reveal otherwise hidden words in multiword terms. An excerpt from the index to the PAIS *Subject headings list* is as follows:

	Industrial relations
Boycott	(industrial relations)
	Industrial relations consultants
Grievance procedures	(industrial relations)
	Industrial safety
Social service,	Industrial
Sociology,	Industrial
Spies,	Industrial
	Industrial surveys

Although the word 'industrial' would be at the front of many of these entries, i.e. the access point, in others it would not be found at all easily by any means other than this index. An interesting use of KWIC techniques has been introduced by DIALOG.[3] Many of the files now contain several million references, and even a well-planned search may well retrieve an unacceptably large number of references from a file this big. To facilitate looking through the results of such a search, it is possible to display titles (or in some cases text) containing the sought term in context. This should be of help in pursuing the search further to improve the relevance performance.

Keyword out of context (KWOC) indexing

Because the filing word is not in the usual place, KWIC indexing looks unfamiliar, and another method of title manipulation is to have the keyword at the beginning of the line, followed by the complete title. This has the advantage of having as familiar appearance – filing word at the left – and also of presenting the whole title as it stands, but is not as successful as KWIC in bringing together titles which contain the same pairs of words. The point is well illustrated by the British Library

Document Supply Centre's *Index of conference proceedings*; this started out in 1965 with a single-word KWOC index, but has had to develop a more sophisticated system using pairs of words as its collection has grown to some 18,000 titles each year. For example, 'A proposed new structure for food and agricultural policy' is indexed under Food policy; agricultural policy; AAAS [organizer], but not under Policy. 'Listening devices and citizens' rights: police powers and electronic surveillance' is indexed under the pairs of words Listening devices; Citizens' rights; Police powers; and Electronic surveillance. There does not appear to any way to get to this starting at the single word 'Rights'.

Using KWIC and KWOC, each title will give rise to a number of entries: as many as there are significant words in the title. For this reason, they are usually used as indexes, leading to descriptions of the documents in a separate file. Judged by the criteria for IR systems in general, they do not perform particularly well. Relevance is certainly likely to be high, in that a title found by looking for a particular word is likely to be useful, but we may have to look through a number of entries at that word before finding a title which looks like what we want. Recall, however, is likely to be low. As we have seen, authors usually look for unique titles, and we have no way of identifying related terms, such as synonyms, other than our own knowledge, which is clearly outside the system. Specificity depends on the authors' choice of words, while exhaustivity again depends on how detailed the titles are. Despite this potential disadvantage, KWOC title entries were found to be popular with users in the Bath University Comparative Cataloguing Study,[4] and must obviously be taken seriously.

Since the 1960s, authors of scientific and technical papers have been encouraged by professional societies and the US Government to give their works explicit titles. Although printed indexes based on titles are not common now, titles are an important source of terms for computer database searching in science and technology, though we should perhaps except patents from this; their titles are required to be accurate statements of the subject area of the contents, but are often made unhelpful in order to avoid helping competitors. In the social sciences and humanities there are also problems of terminology which make title-based indexing less useful.

Citation indexing

Documents of value are likely to contain bibliographies; this is the way in which authors show the foundations on which they have built. Garvey suggests that the list of references is a key part of any scientific paper, since it helps to put the research into its proper context in the development of the scientific consensus.[5] His research also suggests that the use of scientific literature occurs at two quite distinct stages of a research project. Scientists will probably begin any research work by finding some references, but their major use of the literature may not take place until their work is completed and is being prepared for publication; at this stage, they are trying to show the relevance of their work to what has gone before, and the citations may reflect this concern rather than indicate the sources actually used during the research. This would seem to lend weight to the significance of *citation indexing*.

There is a link between a document and each work cited in its bibliography; we can invert this, and say that there is a link between each item cited and the work citing it. Since documents usually cite several items, by scanning large numbers of original documents we can establish much larger numbers of such links. If we now file these according to the items cited, we shall bring together all the documents which have included a given item in their bibliographies. This is the basic principle of citation indexing.

As with title-based indexes, the use of the computer made possible the practical implementation on a large scale of an idea which had already been in use in certain subject areas, notably legal literature. The Institute for Scientific Information (ISI), established in 1961, now produces various citation indexes, notably *Science citation index* 1961–, *Social science citation index*, 1966–, and *Arts and humanities citation index*, 1977–. Between them, these indexes cover over 5,000 key periodicals; these are scanned, and all the bibliographic links found are entered into a computer. The information thus gathered is used to provide the citation indexes, source indexes and corporate indexes; it is also used for an SDI service, ASCA (Automatic Subject Citation Alert), and a 'subject index', the Permuterm index, which enters each item under pairs of significant words found in the title.[6]

To use a citation index, it is necessary to have the reference for a relevant article, but it is often the case that a search starts from such a basis. If we do not have a 'prompt' article, we can use the Permuterm index to try to find a starting point using pairs of keywords from our search formulation; this can be useful in locating information on subjects commonly described by a word pair such as 'Holy Grail', which may well be used in the title.[7] This may then give us one or more articles to use as our starting point for a citation search. From the citation index we can find brief details of other more recent articles which have cited the one we already know. We can turn to the source index for full details, and then locate the articles in the appropriate periodicals. If they are not relevant we can discard them, but if they are, we can use the articles they cite as the basis for a further search in the citation index. By this process of recycling we can compile a substantial bibliography from our single starting point. It is of course possible to do this manually, but at a very substantial cost in time and effort. A search which takes weeks to do manually may be done in minutes using a citation index.

Since every item in the periodicals scanned is entered, we can follow up amendments and corrections to previously published articles. These often contain important information – for example, the withdrawal of a claim of success! – but are usually ignored by conventional indexing and abstracting services. This advantage is of course not inherent in citation indexing, but is reflected in the cost of SCI and its fellows compared with most conventional services.

Citation links

Two articles which both cite another earlier article must have something in common; if they both cite two earlier articles, the linking is increased. This is known as *bibliographic coupling*, and if two articles have half a dozen articles in common, we should be justified in assuming that they covered very much the same subject. (An article which had all 50 citations in common with another turned out to be a translation!) This is a reflection of the fact that authors normally cite those works which constitute the basis from which they begin their own writing. Bibliographic coupling was shown to give good results in studies carried out at MIT. Another approach which has been shown to give useful results is *co-citation*, i.e. the citation together of two or more articles in more than one paper. For example, when bibliometrics began to be widely studied, it was usual to find S. C. Bradford's book *Documentation* cited, since it was in this work that he first published his ideas on 'Bradford's Law of scattering' to a wide audience. If we look carefully, we find that nearly all the works which cite *Documentation* also cite an article by B. C. Vickery in the *Journal of documentation*, **4** (3), 1948. Even if we did not have the title of this article, we could assume that it related to Bradford's Law because of the pattern of co-citation. As the study of bibliometrics developed, other works such as *Human behavior and the principle of least effort* by G. K. Zipf were also co-cited. The study of such patterns would be both difficult and tedious using manual methods, but the availability of citation indexes makes them relatively simple, and we can follow the development of an idea through various stages.[8]

Derived indexes such as SCI require no intellectual effort at the input stage, since in effect they are based on the assumption that the author has done the necessary work to establish the citation links for us. They assume that authors are familiar with the literature of their subject and will quote the appropriate sources fully and correctly; that they will not indulge in unjustified self-citation, and do not ignore documents which put forward relevant but opposing views while quoting articles of marginal relevance by their friends. All of these assumptions are by and large justified, but it would be unwise to think that authors are not as liable to sins of omission and commission as anyone else. Nevertheless, there is no doubt that these tools are an extremely important addition to the range of bibliographical services available to the information worker. They also are based on a common approach to a search for information, where users start with a document which has aroused their interest and which can be used as the start of a search in a citation index. The following example shows how a journal article is treated in a citation index.

The original appears in the **author (source) index:**

> Johnson, Karl E. 'IEEE conference publications in libraries', *Library resources and technical services, 28* (4) October/December 1984, 308–314.
> [IEEE = Institution of Electrical and Electronic Engineers]

It contains the following references at the end (among others):

> Marjorie Peregoy, 'Only the names have been changed to perplex the innocent', *Title varies* 1:13 (April 1974).
> Jim E. Cole, 'Conference publications: serials or monographs?' *Library resources & technical services* 22:172 (Spring 1978).
> Michlain J. Amir, 'Open letter to IEEE', *Special libraries*, 69:6A (Nov. 1978).
> Michael E. Unsworth, 'Treating IEEE conference publications as serials',

Library resources & technical services 27:221–24 (Apr./June, 1983).

The following entries would appear in the **Citation index:**

Amir, Michlain J. ‘Open letter to IEEE,’ *Special libraries*, 69:6A (Nov. 1978). **Johnson, Karl E. 1984.**

Cole, Jim E. ‘Conference publications: serials or monographs?’ *Library resources & technical services* 22:172 (Spring 1978). **Johnson, Karl E. 1984.**

Peregoy, Margaret. ‘Only the names have been changed to perplex the innocent,’ *Title varies* 1:13 (April 1974). **Johnson, Karl E. 1984.**

Unsworth, Michael E. ‘Treating IEEE conference publications as serials,’ *Library resources & technical services* 27:221–24 (Apr./June, 1983). **Johnson, Karl E. 1984.**

The following entries would appear in a **Permuterm index:**

Conference	IEEE	Johnson, Karl E. 1984.
Conference	Libraries	Johnson, Karl E. 1984.
Conference	Publications	Johnson, Karl E. 1984.
IEEE	Conference	Johnson, Karl E. 1984.
IEEE	Libraries	Johnson, Karl E. 1984.
IEEE	Publications	Johnson, Karl E. 1984.
Libraries	Conference	Johnson, Karl E. 1984.
Libraries	IEEE	Johnson, Karl E. 1984.
Libraries	Publications	Johnson, Karl E. 1984.
Publications	Conference	Johnson, Karl E. 1984.
Publications	IEEE	Johnson, Karl E. 1984.
Publications	Libraries	Johnson, Karl E. 1984.

The above examples do *not* have exactly the same layout as you will find in, say, *Science citation index*, but the principles are the same. Once the data has been entered from the original journal article, all the rest is produced by computer.

Summary

This chapter has dealt with methods of producing printed indexes by computer from information manifest in a document: the title or the bibliographical references. The citation indexes mentioned are now available online in parallel with the printed versions; the choice of which version to buy becomes an economic one depending on the amount of use made of the service. Many of the co-citation studies mentioned are really only practical with the online versions.

References

1 Craven, T. C., *String indexing*, Orlando, Academic Press, 1986. Probably the best text on KWIC/KWOC and similar indexes.

2 Luhn, H. P., *Keyword in context index for technical literature*, IBM, 1959. Included in *Theory of subject analysis . . .*

3 *Chronolog*, 15 (2), 1987, 25, 27 (announcement).

4 Bath University Comparative Catalogue Study, *Final report*, Bath University Library, 1975. 10v in 9. (BLR&DD report 5240–5248).

5 Garvey, W. D., *Communication: the essence of science*, Oxford, Pergamon, 1979.

6 Garfield, E. *Citation indexing: its theory and practice in science, technology and humanities*, New York, Wiley, 1979.

Ellis, P., Hepburn, G. and Oppenheim C., 'Studies on patent citation networks, *Journal of documentation*, **34** (1), 1978, 1–20.

Students should examine at least one of the citation indexes produced by ISI in depth, using it in various ways to test its effectivess; cf Brahmi, F. A., 'Reference use of *Science citation index*', *Medical reference services quarterly*, **4** (1), 1985, 31–38.

7 Mann, T. *Library research models: a guide to classification, cataloging and computers*, New York, NY, Oxford University Press, 1993.

8 Small, H., 'Co-citation in the scientific literature: a new measure of the relationship between two documents', *Journal of the American Society for Information Science*, **24** (4), 1973, 265–9; 'Co-citation context analysis and the structure of paradigms', *Journal of documentation*, 36 (3), 1980, 183–96.

Bichteler, J. and Eaton, E. A. III, 'The combined use of bibliographic coupling and co-citation for document retrieval', *Journal of the American Society for Information Science*, **31** (3), 1980, 278–82.

Broadus, R. N., 'Citation analysis', *Advances in librarianship*, **7**, 1977, 299–335. (The application of citation analysis to library collection building.)

Chapter 4
Developments in information technology

While this book is not a text on information technology, in order to examine the use of computers for information retrieval we must look at some of the rapid developments in computer technology which have taken place during the last 35 years, and more particularly the last 15, as these have major implications, both for present practice and for the future. The techniques described in Chapter 3 were feasible with the technology available c1960, but many of the techniques to be covered later have only become possible with more recent technology. Those familiar with these developments can move direct to Chapter 5.

Computer-controlled typesetting

During the 1960s, computers were large centralized installations, with limited access. The user had little control over the end product, and almost all processing was done in batch mode, not online. Input was through the use of 80-column punched cards;. printing was by high-speed (but low quality) line printers. *Computer-controlled typesetting* was developed at this time, much of the work being done for organizations such as the NLM in its MEDLARS project to computerize the production of *Index medicus*. The widespread adoption of this technology has meant that a high proportion of formally published printed material today is also available in computer-readable form; before this, anything to be processed by computer had to be specially keypunched, an expensive and time-consuming operation which in effect repeated all the work that had gone into the production of the original document. It is for this reason that bibliographic databases do not cover material published before the 1960s, and projects to make historic materials available in computer-readable form have only recently become technologically and economically feasible.

Microcomputers

Developments in semiconductor technology led to the introduction of integrated circuits and the microprocessor, which in turn led in the mid-1970s to the first microcomputers, but these were strictly for the 'amateur', that is, those who enjoyed computing and knew enough about it to be able to program the machines for themselves. It was not until about 1980 that the desktop or *personal computer* (PC) became a practical proposition for general use, and was quickly adopted by business and industry as well as in education. Since then, developments in technology and software mean that anyone can now have as much computing power on their desk as was available in mainframe computers 20 years ago.

With the increase in computing power came the need for increasing memory, both for processing and for storage. The first PCs had 64k (kilobytes: 1k = 1024 (2^{10})) bytes of *Random Access Memory* (RAM) for processing, and used floppy disks for permanent storage. (One 360k floppy disk might store the equivalent of about 100 A4 pages.) New operating system software made it possible to use 640k of RAM, then increasing amounts; the current 'minimum' is 4MB (Megabytes), and most IBM-compatible machines can have up to 32MB. Upper range desktop machines can have several hundred MB of RAM. The first *hard disks*, for large-scale permanent storage of files, held 10MB; such a disk would not now hold the operating system software needed to run the computer, and disks holding several hundred megabytes are now common.

CD-ROM

Hard disks are normally a permanent part of the machine – hence the early name fixed disk – but the development of the laser-read compact disc has given the capacity to store up to 680MB (1995 capacity) of data on a removable CD-ROM (*Compact Disc – Read Only Memory*) disc. CD is the format usually found, though it is not the only kind of laser-read disk. Another format which appeared earlier than CD-ROM was the 12 inch laser disk; this is an analog device, with data recorded in a format compatible with television standards, rather than the computer-readable digital form of CD-ROM. It has been used successfully to store illustrations, for example photographs in archive collections. As the name implies, CD-ROMs are read-only devices, but they can of course be quickly interchanged to give access to a variety of sources of information. The adoption of an international standard (ISO 9660 – also known as the High Sierra standard) led to a rapid increase in the number of databases available on CD-ROM, covering a wide range of information. Standardization meant that discs from any supplier could be read on any CD-ROM drive: the market was no longer limited to those people having the matching equipment.

One CD-ROM can store the whole text of an encyclopedia, complete with illustrations – including video clips – and sound. Perhaps more immediately significant is the possibility of storing whole bibliographic databases, including full text, so that we are no longer constrained by the need to link up to a central computer, but can carry out searches on our own PCs. For many users, this is not only more convenient but also less stressful! (It also avoids telecommunication costs.) Some of the implications of these technological advances will be discussed here, but we cannot hope to cover them in depth; further reading is essential to gain a full appreciation of the possibilities.

Networks

Unless we are able to connect our own computer to other computers, we are restricted to those databases we have on our own machine or on a portable medium such as CD-ROM. The use of external databases requires us to be able to connect computers together. In the early days of interconnection, mainframe computers were made available to multiple users through direct links to ‘dumb’ terminals: dumb, in that they could not operate independently of the mainframe. The development of the microcomputer meant that these could be linked to other computers, but act as intelligent terminals, in other words carry out processing on their own. It was more sensible to utilize the power of the local machine even when connected to a large central computer. In order to do this, two things were necessary: telecommunication links and suitable software.

Software

Local Area Networks (LAN) were developed using software which connected PCs to other PCs as well as to mainframes within the same organisation. At the same time, the idea of connecting computers in distant locations through Wide Area Networks (WAN) was implemented by military, academic and commercial users. In 1969, the US Department of Defense set up a network, ARPANET; its purpose was to prevent the complete dislocation of the military network in the event of a nuclear attack, by distributing the computing power to a number of widely separated sites. In the mid-1980s the National Science Foundation (NSF) saw the possibilities of using this technique to keep down the cost of research involving supercomputers; instead of every academic research centre having its own supercomputer, which would probably be underused, a limited number of centres with supercomputers linked together through NSFNET and accessible to other users would be much more cost-effective. Other academic networks have been established, for example JANET[1] in the UK, NREN in the USA, and AARNET and the Australian Education Network (EdNA) in Australia; as more and more networks were linked together, they became the basis of the *Internet*, the network of networks linking millions of users in countries all over the world. Commercial enterprises such as banks and airlines also quickly saw the value of networked computing services to link branches in the same city, same country, and internationally, and have developed highly sophisticated software to facilitate banking and travel worldwide. Unlike the Internet, such networks, and those developed for military purposes, are not for public use, though some hackers have contested this.

To facilitate the use of information on one computer by another remote computer, the concept of *client–server* software has been developed. The computer supplying the information is the server, and the software is designed to supply information in an appropriate format to other computers, possibly to more than one at once. The computer using the information is the client, which must again have the right software to utilize the information in the format supplied by the server. If we are talking about the transmission of simple text, this is not difficult, and there is little difference between client and server; the standard code for the transmission of text, ASCII, is discussed later. Files may also be transmitted in binary format; this is the format required for multimedia files, which tend to be large. To utilize these files, users must have appropriate software on their PC, in addition to the software required for the network.

Telecommunications

Developments in telecommunication links have been a major factor in the growth of distributed computing. To begin with, the existing telephone network was used; since this used electromechanical equipment, it was subject to fairly high error levels, and transmission rates were slow – originally 110 bps (bits per second), which soon rose to 300bps. (The number of bits per second transmitted is often referred to as the *baud rate*, but at the higher speeds becoming common the two are no longer exactly equivalent.) One problem is that computers produce digital signals consisting of a series of 0s and 1s; standard telephone circuits are analogue devices, which represent sound by a continuously varying electrical voltage. To transform the digital signals into a form which can be transmitted over telephone wires requires a *modem* (*mo*dulator-*dem*odulator) which converts the signals from digital to analogue at one end, and analogue to digital at the other. Not only is there the possibility of noise through the telecommunication process, there is also the possibility of error in the conversion process. In recent years, the old electromechanical devices have been replaced by electronic, and there have been improvements in both hardware and software for modems, so that the present (1995) limit is 28,800 bps, accepted as an international standard (V34) in 1994.

Compared with the speed at which computers now operate, this is still very slow, and considerable efforts are being put into increasing transmission speeds. One technique involves making more efficient use of the existing telephone network, which represents a massive investment in infrastructure. *Packet switching*, introduced in the 1970s, meant that it became possible to link networks over long distances at reasonable cost. A normal telephone call monopolizes a certain proportion of the available links between the two centres involved – the ‘bandwidth’ – but the amount of information transmitted only occupies a limited part of the time. In between words, for example, nothing is being sent in either direction. Computers could get through a considerable amount of processing during these dead periods, just as they can between keystrokes. Packet switching takes the input from a number of messages and divides it up into labelled compressed ‘packets’; the packets are then sent in a continuous stream to the receiving station, where they are sorted out into the original messages, which are forwarded to the intended recipients. Dead time is eliminated, and much more traffic can be sent along the same transmission channel using the X.25 protocol.

In the mid-1980s the suggestion was put forward that the telephone network could be used to transmit digital signals in an *Integrated Services Digital Network* (ISDN). Such a network could not only carry telephone conversations and link computers, but would also provide the means for the delivery of video signals, at a speed of 128kbps or higher. At the moment, the change to ISDN is still slow, but it seems likely that new international standards and rapid falls in cost will soon change this. With an ISDN link, there is of course no need for a modem at either end for computers to be able to communicate, since the digital output and input can be transmitted directly without having to be converted to analogue form.

Conventional telephone lines consist of pairs of copper wires (known as 'twisted pairs'); the amount of information that they can carry is quite limited. A significant development has been the introduction of *fibre-optic cable*. Electrical signals are transformed into pulses of light and transmitted down a long, very thin glass fibre cable; at the far end they are transformed back to electrical signals. Transmission can be at very high speeds: currently, 50M bps is usual, while it seems likely that speeds of 2G (10^9) bps will be possible soon. In developed countries, inter-city telephone links are now being replaced by fibre-optic cables. The massive increase in carrying capacity means that one cable can carry the same number of signals as a very large number of twisted-pair copper wire links. Fibre-optic cables are a practical means to carry television signals as well as voice, for example. The growing network of fibre-optic cables is the main structure of the 'information superhighway' of which we hear so much. Just as a freeway can carry more traffic at higher speeds than a network of local roads, so the fibre-optic network can carry more information than the conventional telephone network.

A third type of link uses microwave transmission. This needs a dish aerial at each end of the link, but does not need other connections. This method is being used for the transmission of data in the same way as telephone lines. The limitation is that there has to be line-of-sight transmission from one point to the next, but if this can be achieved, costs are comparable with cable. The use of satellites means that microwave (e.g. TV) transmissions can also be broadcast, i.e. distributed worldwide, and dish aerial receivers for this purpose are becoming a common sight (information from the sky!). Technical problems in linking computers into networks are now largely solved, though the search for increased speeds of transmission will surely continue. We thus have at our disposal the physical means to access information held in a wide range of computers. We should now consider how this information is presented.

Graphical user interface

A significant development in the microcomputer world was the introduction of the Apple Macintosh in 1984. Prior to this, computer commands had had to be typed in, either from memory or by consulting a manual. For many users, this was a tedious and error-prone activity, which tended to restrict the use of PCs to those with the necessary skills. The Apple Macintosh was the first microcomputer to use a *graphical user interface*, GUI, in which commands were represented on the screen by icons, which were selected by a 'pointing' device such as a mouse. The mouse was used to move the cursor to the desired position on the screen; one or two clicks of the mouse button(s) then activated the required command. Alternatively, menus – lists of commands – could be selected in the same way. The typing of commands was reduced to a minimum, making the machine much simpler to use. In addition, the use of graphics within programs was made easier by the fact that the whole screen display was a graphic; inserting illustrations required less complex programming than for text-oriented screens. A GUI of equivalent quality for IBM-compatible PCs had to wait until Windows 3.1 in 1991.

Input devices

For text, the *keyboard* is still the standard means of input. Keyboards originally looked very much like a typewriter keyboard with some extra keys. Despite some changes, the standard shape is so well established that it seems likely to persist, just as the typewriter keyboard layout has persisted. (It is a sobering thought that the QWERTY layout was intended to slow typists down so that they could not type faster than the mechanics of the typewriter could operate!) A device which has proved its use in converting text to digital form is the *scanner*, which will also convert graphics, including those in colour. This has meant that whole libraries of graphics are available for incorporation into documents as needed. More recent is the use of digital video cameras, which enable us to add photographs or video clips to our documents, again in colour. Another device which will certainly grow in importance is voice input; a multimedia PC already has the facility for voice output from text, but voice input is also becoming practical, though as yet it is still basically experimental.[2]

Desktop publishing

The introduction of the laser printer in 1985 meant that computer output was no longer restricted by the limitations of less sophisticated printers. A variety of fonts could be used, with high quality graphics incorporated into the text, and software was quickly developed to take advantage of these possibilities, including printing in colour. Many people now produce their own brochures and pamphlets. Although all of these are computer-produced, by no means all go through the normal publication channels to be picked up by a bibliographic agency; those which do not become part of the rapidly increasing ‘grey’ literature,[3] and may only be tracked down by accident. The other side of the coin is that useful works, such as some specialized publications which only warrant a small edition, can be produced in this way and printed out on demand. Modern high speed laser printers and binding machines enable a bookshop or desktop publisher to hold the text in computer-readable form, and produce a copy ‘while-you-wait’. Computer-controlled typesetting made most formally published material available in computer-readable form; desktop publishing has done the same for informal publications.[4]

Electronic publishing

One important result is the development of *electronic publishing*. Some materials, including journals, are now being published online, with distribution via the Internet.[5] Since there have been in effect no means of charging for material made available in this way, it has been suitable for publications such as academic journals, weather reports and government documents which are not intended to make a profit. With the commercialization of the Internet, charging is becoming practical, and many other journals are now becoming available in this way.

The paperless office

In many organizations now each staff member has a desktop computer forming part of the local area network (LAN). Information does not have to be distributed in the form of print on paper, but can be circulated by electronic means using suitable software.[6] Individual users can build up their own personal files of information useful to them, and information can be circulated using electronic mail to one person, to a selected group, to the organization as a whole, or to anyone who wants to read it. (Security procedures have to be in place to ensure that information only reaches those entitled to read it!) Large organizations such as ICI and the CIA were the first to adopt this method of working, but so far it has not been the panacea that was once assumed, and most offices are still paper-based. (In some cases, there are legal requirements which require documents to be on paper.) It is possible to transmit fax messages from computer to computer, but the fax machine with its paper copies is still the norm, even though most of the messages are produced by word processors and printed out for transmission.

Bulletin boards have been accessible for some years; these consist of computers accessible by modem and telephone line. Some are the work of enthusiastic amateurs, many are established by organizations of one kind or another. Messages may be put on the board by the SysOp (SYStem OPerator) who manages the BBS, or users who connect to the board from their own PC. Most are freely available, including some which are major sources of public domain or shareware programs, but others, e.g. Compuserve, charge fees but provide a wider range of information and services. The distribution of information via electronic channels is now pervasive.

HyperText

If we are reading a novel, we expect to read it straight through, in order to follow the development of the plot and the characters. If by contrast we look something up in an encyclopedia, we may well wish to follow up a reference to an article elsewhere in the work, perhaps in another volume; from there, we may well be referred on again to yet another article. Trains of thought are not linear, as was pointed out in a key article by Vannevar Bush.[7] The idea that this kind of browsing from prompt to prompt could be done by computer was first suggested by Ted Nelson,[8] but the first practical application on a wide scale came with the introduction of Hypercard on the Apple Macintosh. This enabled the user to compile a file (stack) with built-in links outside the main sequence. The name given to this form of file structure by Nelson was *hypertext*, from the idea of multi-dimensional hyperspace put forward as a mathematical concept in the 19th century. HyperText is now generally available and is widely used in databases on CD-ROM, for instance. (It is interesting that both Bush, with the Memex machine, and Nelson, with XANADU, envisaged processing by a very large central machine, not the distributed processing that has now become the norm.)

HyperText links begin with the starting point of the link, an *anchor*. It is then necessary to specify the exact location to which this is to be linked, making a hyperlink to a second *anchor*. Once this is done, the new location can in turn become a starting point anchor, with further links being generated as needed. These may be specific, taking the user to a specific location; local, taking the user to any chosen point in the current document; or generic, taking the user to any point in any document – which may reside on a totally separate computer, which may in turn lead on to a file on yet another computer. It is often easy to lose track of where one has got to!

Multimedia and hypermedia

A GUI is able to display graphics as well as text. With the growth in power of computers, sufficient memory became available to make this feasible at acceptable cost; graphics take up much more disk space than text files. Graphics also require more processing power than text. It was not until 32-bit processing rather than 8-bit or 16-bit was developed that graphics became a practical proposition, with the Apple Macintosh, followed some years later by the Intel 80386 and later processors for IBM-type PCs. Desktop computers to be used specifically as graphics workstations may use 64-bit processing. The change from 8-bit to 32-bit processing was accompanied by increases in the speed at which microprocessors could operate; where an early 8-bit processor might operate at 4MHz (1 Hertz Hz = 1 cycle per second), or 10MHz in turbo mode, 1995 models may operate at up to 130MHz. These high speeds are essential for the processing of graphics, especially video.

The digital recording of sound, first on LP records and then on CDs, meant that sound too could be incorporated into computer files. Sound files also occupy space: the chord which introduces Windows occupies about 25k for a second or so of playing time. The introduction of graphics and sound into computer files gave rise to *multimedia* presentations, which may now also include video and animations. From this, the next step was *hypermedia*: multimedia in which we can jump from one point to another via hypertext links.[9]

Interchange of data

One of the most important limitations of the text-oriented screen is the fact that it can only display a limited number of symbols: upper and lower case letters, numbers and punctuation. Various codes were developed to represent these symbols in binary code, but the one now commonly accepted is the ASCII (American Standard Code for Information Interchange) character set.[10] The sequence of codes is important, because it determines the filing order of the various symbols. The use of ASCII codes may thus dictate the filing order found in computer-based lists of subject headings.

This standard is used very widely, but it caters only for the roman alphabet, and not for any of the accents and special characters used in European languages. An extended ASCII 8-bit set exists, but this does not cover all of the requirements conveniently even for European languages. It is evident that the exchange of information is severely limited if we are restricted to the ASCII character set, yet this is still the standard code used for e-mail, simply because it is standard! There is no guarantee that the sender and receiver of an e-mail message are using the same extended symbols, but if they restrict their text to standard 7-bit ASCII it will be received as sent. Languages such as Greek which do not use the roman alphabet may be forced to adopt complex combinations of ASCII codes.

One severe limitation for anything but the most simple message is that no information can be transferred about formatting. Word processors enable us to use a variety of fonts, font sizes, emphasis (**bold**, *italic*, underline) and layout, to produce documents which look good and convey our meaning effectively. None of this can be transmitted using standard ASCII. Before the introduction of the GUI, this was not vital, as the effects could not be displayed anyway, except indirectly; with a GUI display, all these special effects *can* be shown on the screen – but they still cannot be transmitted to other computers using standard ASCII codes. It is possible to encode them in binary form and transmit this – most communications programs will transmit binary files – but they will make no sense at the receiving end unless exactly the same word processor is used to display them. On the other hand, the ability to display format on a GUI may lead us to neglect structure and content in favour of layout; Honan[11] argues that for large documents a text-based word processor is just as effective as one based on a GUI, since we are forced to pay attention to content rather than allowing ourselves to become unduly obsessed with presentation.

Standards

Both hardware and software are subject to change, often quite rapid. There was a need for standards to be developed which would make possible the exchange of data regardless of the software and hardware used to produce them and receive them.[12] The first step was the development of SGML: Standardized General Markup Language, in 1986.[13] This works by tagging each unit of a document, e.g. heading, title, text, so that it can be recognized as such by the receiver. Information concerning layout and type faces can also be encoded. The widely used HTML: HyperText Markup Language is a subset of SGML. It provides tags for headings, title, address and so on. Each tag must begin and end. The net result is a document which uses standard ASCII codes and can therefore be transmitted simply to another computer, where it can be decoded by suitable browser software. A simple page might look something like this:

```
<html>
<head>
    <title>Welcome</title>                              [headings]
    <h1>Welcome!</h1>
</head>
<body>
                [body tags and text]
</body>
</html>
```

Each tag begins and ends, so that the receiver can recognize the various parts of the document. The end is shown by preceding the tag with a slash /.[14] For those reluctant to undertake the coding task, there is software which will do it automatically, or (at the receiving end) strip off the codes and give plain ASCII text.

A number of other standards have been developed including Open Document Architecture (ODA) and Standard Page Description Language (SPDL). One significant venture was the Text Encoding Initiative (TEI), which was developed in the UK to permit the exchange of documents between universities.[15] To send text as an ASCII file by e-mail would have been fast, but this would have been negated by the amount of work involved in reconstituting the document complete with formatting. The TEI used a form of SGML to solve the problem. Another possible solution is to use the coding/decoding programs UUENCODE.EXE and UUDECODE.EXE; these convert computer files into ASCII strings which can be transmitted over the Internet and decoded at the receiving end.

The introduction of multimedia and hypermedia meant that further standards have had to be developed, not just to cover each format separately but also their use together, since they have to be synchronized (*in sync*). It can be disconcerting when the sound and picture are not synchronized! One of the key standards is HyTime, an SGML-based Hypermedia/Time-based Structuring language; a master encoded document serves as the hub for text files, sound files and graphics, linking them all and ensuring that data from each is used at the appropriate moment.

Another important aspect of file transmission is the amount of information to be transmitted. A computer monitor displays information as pixels, single dots of colour; a common resolution of 640 x 480 pixels contains 307,200 pixels. For 256-colour displays, each pixel requires 8 bits (28 = 256), giving a total of 2,457,600 bits of information, or 307,200 bytes. Thus a static picture in 256 colours occupies over 300kB of disk space. The Macintosh uses 16-bit colour, giving 65,536 colours, requiring twice as much space. To show moving images, it is necessary to repeat the picture 60 times a second, to take advantage of the phenomenon of persistence of vision within the parameters of the computer screen. The amount of information to be stored is obviously very large, and forms of compression have been developed to reduce this to an acceptable level. Here too standards are very important; to use a graphics file which has been compressed we must have the right software to decompress it. If we are transmitting a graphics file through a network using a modem, it is clear that even at 28.8kbps, it will take some time to transmit a file of several hundred kilobytes, possibly some minutes. The most important compression standard is the JPEG (Joint Photographic Experts Group) File Interchange Format; with this, it is possible to reduce a file of 2MB to about 100k, at the cost of some detail.[16] Fortunately the human eye is very tolerant, and the losses are not noticeable. To give some specific figures, the Kodak Photo-CD format gives about one hundred images of high quality on a CD; the Portfolio system gives up to a thousand at lower but still acceptable quality; while compression techniques can give several thousand images on one CD.

We should also not overlook industry standards. Adobe Systems Inc., which developed the Postscript printer control language, has also developed Adobe

Acrobat, software which will take a Postscript printer file and convert it into a form which can be read on any GUI screen in the original format, complete with graphics and colour.[17] Intel have developed the Indeo algorithm for the capture and compression of digital video; this is incorporated in Microsoft's Video for Windows, and Apple's QuickTime and QuickTime for Windows.[18]

The Internet

Over the years, large amounts of information have become available on the Internet,[19] and various programs have been developed to help users find their way about. It must be remembered that there is no overall control of the Internet, no central body to impose order. Any order that exists is the result of cooperation between users. To begin with, most of the traffic on the Internet was e-mail between individuals, but it soon became clear that groups were beginning to form, exchanging the same information among a number of people. This led to the establishment of *newsgroups*; mail sent to the group is automatically forwarded to all the members. There are now several thousand newsgroups around the world, each with its own listserver who manages the mechanics of the subscriber list, and usually keeps track of what goes on to remove 'unsuitable' messages. (Unsuitable may simply mean out of scope – all messages take up space on the server's hard disk –but some may offend subscribers. Walls are not the only place where one finds graffiti!)

To use the network, it is necessary to have software which conforms to the standard TCP/IP (Transmission Control Protocol/Internet Protocol). Telnet allows remote login to other sites to see material that is there. FTP (File Transfer Protocol) enables us to transfer files between our computer and others. To find one's way around we may use Gopher software. There are several hundred servers containing hierarchical menus leading to information available; many library OPACS are accessible using Telnet or Gopher, including the British Library and the Library of Congress.[20] With Gopher client software it is possible to create 'bookmarks' to identify sites that one may wish to visit again; this can save a great deal of typing! Other software such as ARCHIE and Veronica acts as means of locating servers or files.

Though TCP/IP is currently the *de facto* standard protocol for information interchange, there are problems of compatibility with ISDN, and also with the new international standard for Open Systems Interconnection (OSI). Though IBM announced in 1988 that it would begin offering OSI protocol products in 1990, OSI, the *de jure* standard, has not yet replaced the earlier standard,[21] and it may well not do so, in view of the investment in TCP/IP.

The World Wide Web

The difficulties of finding information on the Internet led to the development of a new protocol for linking to computer sites, HTTP: HyperText Transport Protocol. From an idea in 1989, this led to implementation of the World Wide Web at CERN, the European Centre for High-Energy Physics, in 1991. Full-scale operation came in 1993 with the development of the Mosaic Web Browser software by the NCSA

(National Center for Supercomputer Applications), which placed it in the public domain, so that anyone could obtain it using FTP.[22] The Web accounted for 0.1% of NSF Internet backbone traffic in March 1993, after the introduction of Mosaic; by September 1993 it accounted for 1% and by November 1994 10%; use is obviously continuing to grow very rapidly. What led to this sudden ‘explosion’ in the use of the Internet?

The first factor was the use of hypertext to build links between documents. The physicists at CERN were experiencing information overload, and needed a better way to keep track of the publications on the Internet that they found useful. The second factor was the use of multimedia; the Internet had been restricted to text, but the Web software made it possible to use graphics and the facilities of a GUI. Mosaic has now been replaced by Netscape, a graphics browser which enables the user to use point and click techniques to go to other sites and also to create hypertext links using bookmarks. (Other software can also be found from various suppliers.) Each site is identified by a Universal Resource Locator, URL, which may include not only the computer location but also the directory path to specific files.

Another factor in the growth of use has been the interest shown by commercial vendors. While the Internet was largely restricted to the exchange of text between academic institutions, there was little interest from those who did not already have access. With the development of the World Wide Web, demonstrating that graphics and sound could also be used, much more interest was aroused, for example from schools, and various firms have started to offer access on a fee-paying basis. (Although communication on the Internet had always been seen as free, or ‘for the cost of a local call’, the costs had been met by universities, governments and government funded bodies such as the NSF.) As use has increased, so have costs; the availability of graphics and sound meant that much more information is now being transferred between sites, with increased demands on telecommunication facilities. The link between Australia and North America was upgraded in 1995 to double its previous capacity; it took about a day for the additional capacity to be fully utilized!

As was mentioned earlier, there is no governing body for the Internet, nor is there for the World Wide Web. There is also very little control over the information available on it, nor on the way that the information is organized. It can take a great deal of skill to locate all the sources of information of value in a particular subject area, as shown by Westerman;[23] seven business librarians collaborated at length to identify sources of business information on the Net, to provide their users with a service which was small and focused in relation to the Net as a whole. The possibility of using the BSO (described in Chapter 20) to help organize the Net has been suggested; this classification was devised to identify institutions by their overall subject coverage, and might perhaps be used to label sources of information. Whether information providers on the Net would want to cooperate in such a way remains to be seen. Perhaps the increasing presence of commercial vendors will lead to closer control. Users paying for a service are more likely to demand ease of use than those who have access free!

References

1 McClure, C. R. *et al.*, ‘Toward a virtual library: Internet and National Research and Education Network’, *Bowker annual: library and booktrade annual*, 1993, 25–45.

McClure, C. R. *et al., The National Research and Education Network: research and policy perspectives*, Norwood, NJ, Ablex, 1991.

MacColl, J. A., ‘Library applications of a wide area network: promoting JANET to UK academic libraries’, *Information services and use,* **10** (3), 1990, 157–68.

2 Cawkell, A. E., ‘The annual ‘arrival’ of speech recognition’, *Information services and use,* **10** (3), 1990, 133–4. (Editorial) Cawkell’s scepticism is still justified, though progress is certainly being made.

3 Auger, C. P., *Information sources in grey literature*, 3rd edn, London, Bowker-Saur, 1994.

4 Yasui, H., *Desktop publishing: technology and design*, Chicago, Science Research Associates, 1989. This is one of the many books now available on DTP. Students should use a text which is conveniently available.

5 *Infotrain* is an electronic journal produced by students of librarianship, available at http://infotrain.magill.unisa.edu.au

6 Lancaster, F. W., *Toward paperless information systems*, New York, NY, Academic press, 1978.

7 Bush, V., ‘As we may think’, *Atlantic monthly,* **176** (1), July 1945, 101–8.

8 Nelson, T. H., *Computer lib: dream machines*, Redmond, WA, Tempus Books of Microsoft Press, 1987. This text is also available on the XANADU experimental machine.

9 ‘Perspectives on the human-computer interface’ [special issue], *Journal of the American Society for Information Science,* **43** (2), 1992, 153–201.

10 American standard code for information exchange, American National Standards Institute X3.4: 1977.

11 Honan, J. ‘Highway more than a home shopping guide’, *The Australian,* June 20 1995. (Argues very strongly for the importance of text as opposed to graphics.)

12 ‘Workshop on hypermedia and hypertext processing’, *Information services and use, 13*, 1993, 81–199. The need for standards is emphasized by G. Stephenson, ‘Introduction’, 85–7, and by M. Bryan, ‘Standards for text and hypermedia processing’, 93–102.

13 Stern, D., ‘SGML documents: a better system for communicating knowledge’, *Special libraries,* **86** (2), Spring 1995, 117–24.

14 Pfaffenberger, B., *World wide web bible*, New York, NY, MIS Press, 1995. Chapter 27: ‘A quick introduction to HTML’, 447–70.

15 Popham, M., ‘Use of SGML and HyTime in UK universities’, ref. 10 above, 103–9.

Burnard, L., ‘Rolling your own with TEI’, ref. 10 above, 141–54.

16 Bryan, M. In ref. 12 above.

17 Fox, E. A. *et al.* ‘Digital libraries’, *Communications of the ACM,* **38** (4), April 1995, 23–8. (Introduction to a special issue on digital libraries, 23–109)

18 Pring, I. ‘Video standards and the end user’, *Information services and use,* **13**, 1993, 93–102.

19 Krol, E., *The whole Internet users’ guide and catalog*, 2nd edn, Sebastopol, CA, O’Reilly and Associates, 1994. There are a number of good books on the Internet, but this is one of the best and most complete.
Lynch, C. and Preston, C., ‘Internet access to information resources’ *Annual review of information science and technology,* **25**, 1990, 263–312

20 For the British Library, gopher portico.bl.uk. For the Library of Congress, telnet marvel.loc.gov, login as marvel. (Marvel is the LoC gopher.) To use the Library of Congress catalogue, telnet locis.loc.gov and follow the menus.

21 Cawkell, A. E., ‘Videoconferencing, the Information Superhighway and the second Défi’, *Information services and use,* **15** (2), 1995, 73–4. (In *Le défi Americain,* J. J. Servan-Schreiber argues the decline of Europe in the face of American Cultural imperialism.)

22 Books on the World Wide Web, of which ref 14 above is one example, are forming a publication explosion of their own. Many come with floppy disks or CD-ROM, containing software to enable users to set up their own home page. Not all home pages are of value.

23 Westerman, M., ‘Business sources on the Net: a virtual library product’, *Special libraries,* **85** (4), Fall 1994, 264–9.

24 CRG minutes, February 24 1995.

In order to keep pace with changes it is important to scan the computer section of a quality newspaper, and also read widely in the periodical literature; the main problem is not to become bogged down in trivia!

Appendix

Some of the relevant ISO standards are as follows:

ISO 7498:1988 Open systems interconnection reference model. (OSI)
ISO 8613:1989 Information processing – text and office systems – Office Document Architecture (ODA)
Part 1 Introduction and general principles
Part 2 Document structures
Part 4 Document profile
Part 5 Office Document Interchange Format (ODIF)
Part 6 Character content architecture
Part 7 Raster graphics content architectures
Part 8 Geometric graphics content architectures
Part 9 Audio Content Architecture
Part 10 Formal specifications
ISO 8879:1986 Standard Generalized Markup Language.
ISO 8879: 1988 SGML Supplement 1.

ISO 9069: 1988 SGML support facilities: SGML document interchange format (SDIF)

ISO 9541–1: 1991 Font information exchange: Part 1: Architecture.

ISO 9541–2: 1991 Font information exchange: Part 2: Interchange format.

ISO 9660:1987 Volume and file structure of CD-ROM.

ISO 10180: 1993 Standard Page Description Language (SPDL)

ISO 10744:1992 HyTime Hypermedia/Time-based structuring language.

ISO 10918 Joint Picture Experts Group (JPEG) – compression encoding for continuous tone pictures.

ISO 11172: 1993 Moving Picture Experts Group – digital moving picture compression method.

Chapter 5

Derived indexing 2: Database access systems

Background

In Chapter 1, we noted briefly that there has been a dichotomy between bibliographical control systems used for books and those used for other materials. The catchword systems described in Chapter 3 were originally used for book catalogues, but the use of computers made them applicable to many other materials – for example periodical articles, technical reports and conference proceedings. Citation indexing, as developed by Garfield, was specifically applied to periodical articles; books and other sources such as patents may appear in the citation index as items cited (the two most frequently cited authors are the Bible and Shakespeare!), but never in the source index. We do in fact find that there are now two rather different sorts of computer bibliographic databases, those dealing with what might be called macro-publications – books – and those covering micro-publications – periodical articles and all the other similar forms of publication. A practical distinction is that books can stand on library shelves, and can therefore be arranged in some kind of helpful order which is a significant factor in subject searching. The other materials cannot be arranged in this way. It is possible to arrange periodicals as a whole on the shelves, or conference proceedings in book form, but this does not give direct access to the individual articles within them.

The real world is of course grey, not black and white, and this dichotomy is a simplification, but this practical distinction is paralleled by the way that the two streams are treated for information retrieval. Books are 'catalogued' while other items are 'indexed'. Both techniques have the same general objectives: to identify the item and provide access to it through various approaches, including the subject. However, the cataloguing of books usually involves summarization: we treat the contents as a whole, and provide subject access on a limited scale – a class number for shelf arrangement, and one or two subject headings for access through the catalogue. The indexing of other materials tends to be more detailed: we do not have a class number for shelf arrangement, but perhaps for arrangement in a printed bibliography, and we tend to be rather more generous in the provision of terms for subject access. To give a practical example, the 5th International Study Conference on Classification Research, which is cited in several chapters in this book,[1] can be catalogued as a book, shelved at 025.4, and given the subject heading Books—

Classification. In an abstracting journal such as *Library and information science abstracts*, or an indexing journal such as *Library literature*, we would expect find entries for each of the individual chapters, on DDC, UDC, LCC, thesaurus construction, reclassification and so on. While we may regret this separation, it has existed since the beginning of indexes to periodicals, and has practical consequences for information retrieval. In this section we shall be looking at databases covering micro-publications, postponing the examination of book catalogues – manual and OPACS – to later chapters.

General publications

In Chapter 4, we referred to the development of computer-controlled typesetting in the early 1960s. To begin with, this was relatively rare, but three significant groups quickly saw the advantages. Bibliographic databases such as *Index medicus* could be produced more quickly than by conventional methods, and could also be cumulated progressively into a large database which could be searched as a whole.[2] This contrasted favourably with the tedium of a search through a large number of separate printed issues. Since the first tentative experiments in the late 1960s showed that online access was practical, increasing numbers of bibliographic databases have become available online, and over 450 were accessible through the DIALOG Information Retrieval Service, the largest of the utilities of this kind, in 1994, yet this is only a small proportion of the total now available.[3]

The second group to profit from computerization were newspaper publishers. Journalists could now produce their copy on a microcomputer for direct input into the main computer, which can be used to typeset the whole newspaper without the intermediate step of having it composed on a Linotype machine. In fact, this machine, which had been the core of newspaper production for the whole of this century, disappeared from use in a surprisingly short time. In due course, newspapers became available online, and many are now accessible in this way.

The third group were governments, and Statutes and other legal documents began to be produced and made available online with increasing frequency, until this may now be said to be the norm, certainly in developed countries. Legal databases were among the earliest full-text sources to be available online.[4] Commercial publishers also use computer-controlled typesetting, but are naturally reluctant to make their publications freely available! Unlike governments, commercial publishers need to make a profit from the sale of their products. The important question of intellectual property and copyright is looked at in Chapter 28.

All in all, very large amounts of online information are now accessible through a variety of sources, some free but some on a commercial basis. When online databases first became available about 1970, users were mainly librarians and other information workers. The easy accessibility of databases today, and the (relatively) user-friendly software at hand to use them, has meant that many other people have begun to use them without the help of intermediaries. There are however certain skills which must be acquired to make the best use of the information available to us. As pointed out in Chapter 2, not all users have the same requirements for recall and relevance, and search strategies can be modified to recognize this. We must be aware of which databases are likely to prove most profitable in searching for particular kinds of information, and we may have to put some effort into presenting the information that users want in a format that will meet their needs.

DIALOG, the major online utility, was introduced in the 1960s, and was used by both intermediaries – librarians and other information workers, and by end users – users who originated the requests for information. By 1982 the numbers of intermediaries and end users searching the facility was about equal, but by 1988 end users formed about 65% of new users.[5] However, not all enquirers wish to do their own searching; one study at the University of North Carolina showed that a number of users requested help, even though various databases (ERIC, Books in print and Ulrich's Plus) were available on CD-ROM, and BRS/After Dark (specifically intended for end users) was available online. Of the reasons given for preferring an intermediary, over two thirds cited lack of search expertise and over 40% did not wish to spend their own time. Users may also use both options; many of those surveyed planned to use an intermediary again, but also to do their own searches on occasion.[6]

In a work such as this, we can only consider the basic techniques, but textbooks such as that by R. J. Hartley *et al.*[7] go into detail on searching techniques in a way not possible here, where we can only attempt to summarize the main techniques currently in use. Since databases may be online through commercial utilities such as DIALOG, BRS/Search or ESA/IRS, but may also be available on CD-ROM or through the Internet, in line with Dalrymple and Roderer[8] we have used the title database access rather than online searching.

Computer searching of text

If we knew that some information we wanted was to be found in a particular document, but there was no contents list or index, we could find what we wanted by reading through the whole of the document, looking for the words we were interested in. To use the term that we have used previously, we would be trying to *match* our requirements, as expressed in certain words, against the words to be found in the document. Now we may be prepared to do this for one document, but when we start to think in terms of a collection of documents the process clearly becomes impractical. Even to look through a lengthy list of titles can be time-consuming, as many readers have discovered from scanning booksellers' lists and similar publications.

The computer can perform this kind of matching operation at high speed; if the titles or other parts of the text are in machine-readable form, as is now usual, we can program the computer to carry out the matching process and identify the documents likely to be useful to us. All that we have to do is to feed in to the computer the words that we want it to match. With access to large quantities of text either online or on CD-ROM, finding an answer to a particular question becomes a matter of fast, painless extraction. The computer does not become tired of searching – it has no futility point! – nor does it get bored. It simply goes on searching for exactly what we have said we want until the search is completed. (We may of course discover at this stage that what we said we wanted was not really an accurate statement of our needs! Search statements often contain errors, while many are not followed through to obtain a ‘best’ result.)

Early information retrieval systems ran in batch mode, and used magnetic tape as the storage medium. This meant that searching was slow; even a small database of a few thousand items might take several minutes to scan, since the scanning had to be carried out sequentially. (Compare the speed of locating a particular track on an LP or CD with that of searching through a cassette tape!) The introduction of random access storage greatly improved access times, but still required considerable searching to find the items which matched a request. The solution was to create a second *inverted* file in which each term in the original (with the exception of stop words) is listed as an index entry. A search then begins with the index file, which gives the number and locations of postings of a term, and can quickly compare the postings under two terms to determine which are common to both. In some systems, one inverted file lists the terms with the number of postings, and a second lists the specific locations. Basically the same method is used with today’s much larger files. The penalty for the improved access times is that additional index files are required, giving a database perhaps twice as big as the original. Services such as DIALOG require hundreds of gigabytes of storage for their holdings; a good proportion of this is taken up by the index files accompanying the databases. The majority of current databases utilize the inverted file method, but cluster analysis and similar processes are carried out on text as it stands, without the need for inverted files.

Search strategies

We can program the computer to carry out the matching program in two ways, paralleling the changes that have come about in the use of other computer programs such as word processors. The first database programs required the user to type *commands* in a more or less esoteric language, e.g. strs for stringsearch. This meant learning the commands before one could make good use of the service. Since each service used a different command language, this could lead to confusion; fortunately, since all of them perform the same tasks, once one set of commands has been learnt it is not too difficult to master others. Attempts have been made to develop a ‘common command language’, so that all services used the same commands, but this has not had any great success.[9] The alternative to using commands is to use *menus*; these are simpler to use, and do not have to be learnt, but they tend to be slower than using commands direct. (We are only talking of seconds, but the online user prefers to fill the unforgiving minute with sixty seconds’ worth of processing done! The beginner may also wish to make use of Help screens, further slowing down the process, if the menus are not completely self-explanatory.) DIALOG and other services offer the choice for many databases; once users are familiar with the system through using the menus, they can switch to using the commands.

Online searching should then be very simple, but in practice it turns out to be rather more complex.[10] Suppose that we are interested in *classification*; we may find documents using the word *classifying* just as useful. Again, suppose that our interest is in *pollution* –a topic very much in the public mind. We could search for *pollution*, but we might then miss documents which only use the word *pollutant*, or *polluting*. The solution is simple: we can *truncate* our search term to *pollut*. We may need to show specifically that we have used truncation by adding a symbol such as the asterisk *, to show that we are interested in the stem *plus*, otherwise we may find that we get a nil response; some systems automatically take any input as the beginning of a search term, and do not require a specific indicator. We can use forward truncation, e.g. POLLUT*, or backward truncation, e.g. *CLASSIF*; the latter will now match reclassification in addition to words beginning classif . . . However, it will also match *declassification*, the process of making secret documents public. A search for information on the role of the parent could be set up as *PARENT*, to retrieve *parents*, *parental*, *parenthood*, *grandparent* and *grandparents*, but will also retrieve *parenthesis*, *parenteral*, and *transparent*! In some systems, the computer will display a list of terms which will be located by our truncations, which may lead to second thoughts! We may also be able to use a wildcard to allow for variations in spelling, e.g. WOM?N will match women and woman; F?ETUS will match fetus (US spelling) or foetus (British spelling).

It is possible to have a parsing program which will recognize suffixes, and perform the truncation process for us automatically. This is known as *stemming*, and it is important to know whether the software we are using will do this, or if we have to specify a truncation indicator like the * shown above. In the first case, POLLUT will be accepted as a search term, but in the second it will be rejected without the asterisk. Stemming can take place in two stages; the first is at a simple level, and, for example, may convert plural to singular, remove suffixes representing verbal forms (-ing, -ed) and merge variant spellings, (-isation, -ization; -our, -or). The second, more powerful, stage may remove a wide range of suffixes, e.g. -itis, -able. Stemming can be a valuable search aid, but as with manual truncation, it may also lead to disaster![11]

So far we have assumed a search for one word, but in practice we would normally be thinking of more than one word to denote the subject we are interested in. For example, we may be concerned with *water pollution* rather than pollution as a whole. In this situation we can use the ability of the computer to handle logical statements; the logic is Boolean rather than Aristotelian, and it means that we can link the words we are searching for by the operators AND, OR and NOT. Our search becomes:

WATER AND POLLUT*

However, we might remember that water includes sea(s) and river(s), and modify the search to take account of this:

POLLUT* AND (WATER OR SEA* OR RIVER*)

We might want to exclude sewage as a form of pollution:

POLLUT* AND (WATER OR SEA* OR RIVER*) NOT SEWAGE

The order of precedence among the operators is NOT, AND, OR. Thus

POLLUT* AND WATER OR SEA*

will be treated as a search for *pollut** in association with *water*, or *sea** on its own. The use of parentheses in the above example is necessary to avoid this, and is good practice in clarifying a search formulation anyway. It must also be remembered that users unfamiliar with Boolean techniques may use the wrong operator altogether; needing information on, say, 'cats and dogs', they need to specify this as 'cats OR dogs', otherwise they will retrieve only information dealing with both, rather than information dealing with either.

The example above could be set up as a series of simple searches, using the results at each stage as input into later stages, as shown here (system response underlined):

```
1      POLLUT*
Search 1 POSTINGS 732
2      WATER
SEARCH 2 POSTINGS 1653
3      SEA*
SEARCH 3 POSTINGS 451
4      RIVER*
SEARCH 4 POSTINGS 679
5      2 OR 3 OR 4
SEARCH 5 POSTINGS 2215
6      1 AND 5
SEARCH 5 POSTINGS 142
7      SEWAGE
SEARCH 7 POSTINGS 284
8      6 NOT 7
SEARCH 8 POSTINGS 114
```

The operators AND, OR and NOT apply to whole documents; in the above example, we would be looking for documents in which *pollut** was found along with *water* or *sea** or *river**, but *sewage* did not occur. We may wish to be more precise, and specify that words appear in the same paragraph, same sentence, or adjacent, and some services permit this. Note that ADJ may be qualified, e.g. ADJ(n), where n is the separation we are willing to accept. For instance, ADJ(5) means that the two sought terms must occur within five words of each other.

WATER AND POLLUT*	(same document)
WATER SAME POLLUT*	(same paragraph)
WATER WITH POLLUT*	(same sentence)
WATER ADJ POLLUT*	(adjacent)

The order of words may also be important; if we specify INFORMATION ADJ(2) RETRIEV*, we may exclude 'retrieval of information', because the words are not in the specified order, or 'information we want to retrieve', because the words are too far apart. Does the program we are using count stopwords in determining how close words are? We may need to think carefully about the order of words, and the likelihood of their being separated by unsought words, when formulating our search. A possible alternative is to enclose search phrases in quotation marks "..." so that they are treated as a whole. We may search for material on 'circulation' AND 'control' only to find we have retrieved *Control of the peripheral circulation in man*! Proximity searching can also be useful if we are not sure whether we are looking for one word or two, for example post-coordinate or end user.[12] For proximity searching to be possible, the inverted file has to record the location of terms very precisely, so that it becomes much larger than one which simply records that a term occurs *somewhere* in a document.

We may misspell a word used as a search term, either through ignorance or lack of typing skills, or select a word not used; the system may in this case display a list of words close to our spelling in alphabetical order, or sounding the same (assonance), enabling us to correct the error, or choose another search term. (This is very similar to the operation of a spelling check in a word processor.) A powerful method of automatic error correction uses soundex codes; this involves the removal of all other than initial vowels and reduction of the result to four characters. It involves a large dictionary and a great deal of processing, but can cope with most errors other than transposition – which for many unskilled keyboarders is a frequent error![13]

The database may include indexing by a controlled vocabulary as well as text words, that is, terms added by an indexer from a predetermined list. If this is the case, we should be able to display search terms from the vocabulary, and also any related terms, or expand our search to include them; this involves the rather alarming instruction EXPLODE in MEDLINE! In general, this technique is known as the use of *hedges*; a set of terms can be brought together 'within a hedge' to represent a broader subject for which no single term is suitable, by the use of the logical OR. For example, USA OR France OR China OR Russia OR United Kingdom could represent 'the permanent members of the UN Security Council'. Hedges represent the kind of grouping found in classification schemes or similar controlled vocabularies,[14] and are used when searching the text as it stands may produce problems. A search term may have synonyms or near-synonyms, for example drunk driving, drink driving, driving under the influence, drunken driving . . .; or be ambiguous and need context to clarify, e.g. record; or ill-defined, e.g. democracy; or it may occur too frequently to give useful search results.[15] Hedges can be created by the database producers, as in MEDLINE, or by searchers, who may also modify them on the base of experience. They may be based on semantic relationships found in a thesaurus or dictionary, or based on cooccurrence shown by computer processing.

Documents are represented in the computer by *records*; a collection of records of the same kind becomes a *database*; database records are normally organized into *fields*, so that we can restrict our search to one particular area. For example, if we know the author's name we can cut down our search time by searching only the author field, or the title field if we know all or part of the title. If the database includes descriptors from a controlled vocabulary in a descriptor field, we can make use of this in two ways. We can select descriptors from the controlled vocabulary to make our search; alternatively, once we have found a useful document, say by searching on words in the title, we can use the descriptors used to index it to revise our search strategy. Some documents may have an abstract field, or a full text field;

searching the full text of documents online may take time and add to the bill – a good argument for the use of databases on CD-ROM!

If we are carrying out a search involving several words, it is good practice to treat each one as a separate search, to obtain the postings for each, as shown earlier. We can then cut down the search time in the most effective way by beginning with the term having the fewest postings and combining it with the one with the next fewest. We may even reduce the number of hits to an acceptable number without searching on the most frequently used term, which is likely to be the least effective in rejecting unwanted references.

Problems with Boolean searching

All users of Boolean searching quickly become familiar with one of the disadvantages. For a search to be successful, the result must fall within the user's Futility Point; if the FP is m, and a search retrieves n documents, it is successful only if $n \leq m$.[16] A search in a large database on one term frequently gives far too many postings, so we AND a second term; if this still gives too many we AND a third, and so on. Unfortunately, we often find that adding one more term to the search formulation reduces the number of hits from an unacceptably large number to zero.[11] Alternatively, we may begin our search by specifying a number of terms to be ANDed together, only to find that we retrieve nothing at all. How can we best reorganize our search to give an acceptable result? The obvious way in both situations is to drop one or more terms, but which? The tendency is to omit the terms which the searcher thinks are least significant, and retain those which are considered to be most significant. This has been described as the 'anchor effect', and in effect retains certain terms in every variant of the original formulation when their omission might well lead to success.

One way of manipulating the chosen terms is to search on *combinations* of them.[17] For n terms there are $2^n - 1$ combinations, so that if we start with five terms there are 31 possible combinations. Even three terms will give us seven combinations. To work all these out intellectually and try each one would be time-consuming and tedious, but the computer can do it for us, using the kind of effective search strategy mentioned above, and give us the result at each stage. We can thus stop the search while the result is still below our FP, while remaining confident that the search results will match our search closely if not exactly. Some IR systems do offer this facility, known as *quorum searching*, though it is only the more sophisticated software which is likely to have the necessary computing power.

Ranked output

Boolean searching gives us no control over the cutoff point. In other words, we cannot say how many documents we would like to retrieve, and aim for that target; we have to accept what the computer gives us. A more effective approach is to use a more sophisticated search procedure, so that the results of a search can be *ranked* in order of probable relevance. We can then select as many as we wish from the top of the list with some confidence that they will be the most useful. In contrast with

Boolean searching, where the system dictates the cutoff point, this enables *us* to set the cutoff point: we can say that we would like to see the six, or a hundred, or three, most promising documents, and ignore the rest unless we find that they are needed to broaden our search. We are no longer at the mercy of the system, but can set our own parameters for success.

One way to achieve this is to weight the terms used for indexing or searching, but how can we allocate suitable weights? We can either do it intellectually or by means of computer manipulation. In either case we can regard the document-term link as a matrix in which there are *x* documents indexed by *y* terms; w_{ij} is the extent to which term *i* is used to index document *j*.

	Documents										
Terms	w_{11}	w_{12}	w_{13}								
	w_{21}										
	w_{31}										
										w_{ij}	

In searching on unweighted terms, the value in each cell is either 1 (that term *is* used to index that document) or 0 (that term is *not* used to index that document). This is the situation found in Boolean searching.

In a search system using weights, the value of w_{ij} can range from 0 to 1, and may be regarded as the probability that a particular term will be useful in retrieving a document in response to a particular enquiry.[18] To allocate weights intellectually might be possible for a small collection, but would quickly become impractical. We therefore have to consider ways in which weights may be allocated by computer.

One approach is to use statistical methods to indicate the significance of terms. This can be done in a variety of ways. Word frequency counts on their own are somewhat simplistic, but are more powerful if we consider word frequency in relation to expected frequency, based on counts of words in a large body of literature. One method, the Associative Interactive Dictionary (AID) was developed for searching the various MEDLARS files.[19] The inverted file for the database shows us the terms used and the number of postings for each; from this can be calculated the expected frequency of occurrence in any given set of documents of a given term. Let us assume that a search recalls *n* documents from the total *N* in the collection. For any given term which occurs in these documents we can find the total number of postings for that term in the collection *T*. The expected frequency of occurrence *E* is then

$$E = T.n/N$$

If we now calculate the actual number of occurrences *O*, we can derive a relatedness measure (cf the statistical measure χ^2) to show the strength of association between the term being studied and the documents retrieved:

$R = (O - E)/E$

We can then calculate the value of *R* for each of the terms which occur in the set of documents, and rank them in order. A user may enter a search term, and ask to see the related terms; the results are like those found by intellectual effort, but include some unexpected results. One example started with the word 'shellfish' and located a large number of associated terms, of which the nine most highly ranked were:

Rank	XTRA-PSTGS	Term
1	390	OYSTERS
2	334	MUSSELS
3	227	CLAMS
4	185	TIDES
5	180	ESTUARIES
6	143	PARAHAEMOLYTICUS
7	138	CRASSOSTREA
8	101	SEAFOODS
9	88	VIRGINICA

While most of these might be expected, some would not appear in an intellectually derived list! However, they are certainly justified by their occurrence in the documents studied; as Svenonius points out, associations derived from full-text analysis by computer represent the logical extreme of literary warrant![20]

While it is not as powerful as *expected* word frequency, *absolute* word frequency may be of value. Words which occur very frequently in a particular document collection will obviously recall a large number of documents, many of which may be marginally relevant or not relevant at all to a particular enquiry; words which occur infrequently will give lower recall, and will thus enable us to reject unwanted documents more easily. A weighting factor could thus be based on the reciprocal of the frequency of occurrence.

The *term discrimination factor* can be calculated to determine which terms will be most useful in distinguishing one document from another. If we have two documents which are represented by sets of index terms, we can compute a measure of the similarity between them; where the sets of index terms are the same, the similarity measure would be 1, whereas if the two sets had nothing in common, it would be 0. Normally it would lie somewhere between 0 and 1. We can calculate the similarity measures for a collection of documents to arrive at an average figure. We can then recalculate the figure with each term removed in turn; the discrimination factor for each term will be shown by the difference between the similarity measure obtained with that term omitted and the average similarity measure, and terms ranked according to their discrimination factor. Terms which appear very frequently will have a low discrimination factor, and are not good indexing or search terms.

Location may be used to weight terms; words which occur in the title, for example, are likely to be highly relevant to the subject of the document, as are those found in an abstract if one is included.[12] Instead of words as they occur, we may use stemming. In the process of a search, user feedback may be used to weight more heavily those terms which retrieve relevant documents; terms from useful documents found may be added to the search formulation. A combination of search terms can now give a search formulation for each query representing the weighted combination of each of the individual terms.

Many of these methods were tried with success by Salton in the SMART experiments, and are now being incorporated into working systems.[21] Salton proposed that each term in a document should be regarded as a *vector*; the totality of *n* terms would then give an *n*-dimensional vector describing the document. A query would then be treated in the same way, and the two multidimensional vectors matched. One measure which was found useful was the cosine correlation coefficient, which is a measure of the angle between the two vectors; if they coincided perfectly, the angle between them would be 0°; the cosine of 0° is 1. If the two vectors did not match at all, the angle between them would be 90°, with cosine 90° being 0. In practice, a figure between 0 and 1 would be obtained for each document in relation to a query, enabling the documents to be ranked. The function is given by the equation:

$$\cos(\mathbf{q}, \mathbf{d}) = \frac{\sum_{i=1}^{n} \mathbf{d}_i \mathbf{q}_i}{\left(\sum_{i=1}^{n} (\mathbf{d}_i)^2 \cdot \sum_{i=1}^{n} (\mathbf{q}_i)^2\right)^{½}}$$

Alternatively, we may use the rather simpler form:

$$\text{Strength of association} = \frac{Cab}{\surd(Oa^2 \times Ob^2)}$$

where *Oa* is the total number of occurrences of term *a*, *Ob* is the total number of occurrences of term *b*, and *Cab* is the number of cooccurrences of terms *a* and *b*, proposed by Sparck Jones.[27]

One of the systems to use methods similar to SMART is the CITE NLM system,[22] which accepts queries in natural language and uses them as the basis of a search after deleting any stop words (about 600 are used). Using the original words, terms from MeSH, the indexing language used in the descriptor field, related terms derived from computer processing, and the use of combinations of terms as outlined above, the system can give ranked output. The user can use this to make relevance judgements, which can be used to modify the search strategy if the user is not satisfied with the first results.

Maron proposed a rather different approach to calculating the weight to be applied to each term used to index a document. Using the matrix set out above, the weight to be given to a term is the probability that a user requiring a particular document D_j would use term I_i to search for it. This probability w_{ij} can be estimated as:

$$\frac{\text{Number satisfied with } D_j \text{ and using } I_i}{\text{Number satisfied with } D_j}$$

symbolized as $P(I_i|A.D_j)$, where *A* represents the whole set of users. However, this weight is still based on the indexer's estimate, whereas we should be trying to rank documents according to the users' needs $P(D_j|A.I_i)$. We can convert one viewpoint to the other by mathematical manipulation:

$$P(D_j|A.I_i) = P(D_j|A).P(I_i|A.D_j).c = P(D_j|A).w_{ij}.c$$

where $P(D_j|A)$ is the probability that document D_j will meet the needs of all library users A, which is calculated by the kind of statistical techniques outlined above. On this basis it is possible to rank documents according to the probability that they will meet users' needs.[23]

The ranking process does require more computation than Boolean searching, so a possible compromise is to carry out a Boolean search to reduce the number of documents to be ranked to, say, a couple of hundred, and then rank those. The method is now a practical proposition, and is used by some databases, though the majority still use Boolean searching only. The advantages of ranking are such that we shall surely see a steady increase in its use.

Recall and relevance

Most of the work done in establishing the concepts of recall and relevance was carried out on small databases. The 1965 MEDLARS evaluation[24] was the exception, in looking at a database which already contained some 800,000 references. There are now several databases containing millions of references. It becomes apparent that a relevance ratio which might be quite acceptable in a collection of a few hundred references will probably be quite unacceptable in one containing even one million. In the MEDLARS evaluation, the average number of references retrieved for each search was 175, with an average relevance ratio of 50%; that is, of the average 175 references found, about 90 were found to be not relevant. The database is now some ten times as big; following the same search procedures, we would retrieve an average of 1750 references for each search, of which some 900 would not be relevant! The average recall ratio was about 58%, as calculated by a somewhat roundabout method; it was obviously not feasible to examine the whole database in relation to each search in order to establish the recall base. Taking the average search, and assuming that about 90 of the references found were relevant, with a recall ratio of 58% this implies that about 155 references should have been found, but 65 were missed. Again extrapolating this to the current database, we would have to assume that some hundreds of relevant documents would be missed.

We also have the consider another factor in addition to recall and relevance: *utility*, also mentioned in Chapter 2. If we carry out a search using a particular search strategy, we should recall some relevant documents, and we may assume that these will be useful to the enquirer. If we modify the search to recall more documents, we shall almost certainly retrieve some of those found already. To find them a second time is no longer useful! This point was considered in the MEDLARS study, where one of the factors measured was the 'novelty ratio': were the documents retrieved new to the enquirer, or were they already familiar? As was pointed out in Chapter 3, to use a citation index we have to have a starting point which is a document already known to be relevant. It is not a success if our searching eventually reveals the document that we began with – though it is a reassurance that our search strategy is sound. In carrying out a search by any method we will normally arrive at a point where further searching simply re-locates the documents we have already found. It is then time to stop searching, or to adopt a totally new approach!

Blair[16] argues that with the large databases we now have, past thinking on satisfactory levels of recall and relevance is no longer adequate; increases in size have led not merely to a quantitative problem, as illustrated in the previous paragraph, but also to a qualitative change in the way that we must look at retrieval. The ability to reject ('dodge') unwanted material becomes a great deal more significant; we need to achieve much higher relevance ratios, while in certain circumstances much higher recall ratios are essential. The example which Blair quotes in particular is that of a legal database, set up by two lawyers to support their arguments in a court case. Searches were formulated by the two lawyers assisted by two paralegal aides, and carried out by two information specialists. Many of the searches were carried further by manipulating the search formulations, adding terms and using the power of the database program STAIRS. Documents retrieved by the additional searches were passed to the lawyers with those found by the searches that they had formulated. In general, the lawyers felt that they were retrieving about 75% of the relevant documents by their searches, bearing in mind that they had themselves set up the database. In fact, their searches were finding about 20% of the total relevant documents revealed by the various searches. The additional relevant documents were only found by greatly expanded search strategies. In such a situation high recall and high relevance are essential if a case is to be successfully argued. With a large database it is also likely to take much longer to reach the point where we no longer retrieve useful documents. The fact that most large law firms now set up such databases to support their arguments suggests that we should not be complacent about the success at present of online text searching, since search methods are likely to be used a great deal more intensively than has been normal in the past.

Interface design

Vickery and Vickery have a lengthy and very helpful discussion of the overall design of a search interface.[25] This looks at the *functional requirements* of a system, and the *query processing techniques* that can be used to achieve them. Considering the functional requirements first, faced with the need for a search, we must establish the context. This will enable us to select suitable databases and hosts (some databases are available on more than one bibliographic utility, and may also be on CD-ROM). We then have the user's expression of the query, which we may need to clarify through the usual reference interview. We then need to merge or translate the terms used in the query into those likely to be found in the database, for example using a controlled vocabulary, to create a search statement. We can then carry out a search and obtain a set of results. After eliminating duplicate hits, we can evaluate the results, which may be ranked by the system. This procedure is of general application, and applies to manual as well as computer searching.

Among the query processing techniques, we have the need for disambiguation of search terms, possibly using thesaurus relations and classification hierarchies, and the elimination of words in stoplists. We may need to use stemming to remove suffixes. The query may then be formulated as one or more Boolean search statements, which we may need to manipulate to get the best results. Search terms which do not give the desired results may be errors which can be checked against spelling or sound, as mentioned earlier. The system may give us values for term relevance, making document weighting and ranking possible. It should be possible to modify the original query by relevance feedback based on first results.

Computer classification

Full-text databases tend to be large, which means that, as discussed above, they cause problems in use. We may be able use computer processing to help by reducing the amount of work that has to be done to carry out a given search. The similarity coefficients referred to above may be used to give *clusters* of terms or documents to assist in searching.[26] Clusters of terms may be used as hedges, while clusters of documents serve to reduce the bulk of the collection to be searched in response to a given query. To consider the two extremes, we may regard the whole collection (of documents or terms) as one cluster, or we may regard it as consisting of as many clusters as there are documents or terms; obviously neither of these is particularly helpful in processing a search, and we need to find a satisfactory intermediate value.

If we consider terms, we can compute relationships between pairs of terms from the number of times they co-occur in the same document, for example. We can then rank these and set a cut-off point, above which terms may be considered to be related; the cut-off point determines the strength of the relationship. We may exclude terms which occur in only one document, on the grounds that adding such a term can only increase retrieval by that one document; or those which occur very frequently on the grounds that their use would not be helpful. (To take the extreme again, a term which occurred in every document in a collection would have no discrimination value whatever!) Four kinds of group may be found: strings, stars, cliques and clumps. Strings occur when term A is strongly associated with term B, term B with term C, and so on. In practice, strings tend to form loops fairly quickly: term A → B → C → D → E → A. Stars are found when one term is equally strongly related to two or more others. Cliques occur when a set of terms are all strongly related, each to the other. Clumps are a weaker form of clique, in which a term is related to one or more of the others in the clump, but not necessarily to all. In searching, we might begin with a given term but find the results unsatisfactory; we can then use the previously determined relationships between terms to change our search strategy, as mentioned in the earlier discussion of hedges. We normally think of the grouping of terms as a recall device, but Sparck Jones pointed out that this could be a precision device. If we begin with, say, four terms and conduct a Boolean search, we may retrieve nothing; by substituting related terms we may be able to achieve success at the level of coordination that we began with, rather than by simply dropping one or more terms to obtain results at a lower level of coordination.[27]

If we look at clustering from the point of view of documents, we can use the same kind of approach to determine which documents are likely to be related. Instead of using Salton's technique to measure the correlation between documents and queries, we can use it to measure the correlation between documents.[28] We can then form document clusters, within which all the documents will be related to each other at a level we decide. For each cluster we can determine an average 'centre of gravity' (centroid) which represents the cluster as a whole; this may either be a specific document, or a calculated quasi-document. In searching, we can now restrict our processing to the cluster of documents whose centroid most closely corresponds to the query. In practice, by using different correlation level cutoff points, we can build hierarchies of clusters; we can then begin searching at a level which seems most likely to meet our needs, depending on whether we are looking for high precision or high recall. For high precision we would use the clusters with the highest correlation values, which will of course be the smallest; for high recall we might prefer to begin with the larger clusters having lower correlation levels.

For computer-generated clusters to be useful, they must be reasonably *stable*. A method that gives clusters which change significantly each time we add a document will not be particularly helpful. Consistency is also helpful; processing should preferably give one cluster, or at most a limited number. While consistency is not absolutely necessary, in that different clusters may perform equally well in practice, stability appears to be essential. This is most likely to be achieved when the databases which are processed for clustering are large; just as a manually constructed classification scheme or thesaurus will change substantially with each new document classified or indexed while it is still small, it will eventually reach a state where the average change for each document added is relatively insignificant. So it would appear that the databases which are likely to lend themselves best to clustering techniques are those with which it is likely to prove most useful! The first attempts at developing clustering techniques were carried out on small databases; the largest database used with the original SMART experiments contained just over 1000 documents.We now have far greater computing power available to carry out the intensive processing involved in clustering techniques, and we may well see techniques once dismissed as purely experimental become not only practical but also economically viable.

Limitations on computer matching

It is important to remember that computer techniques for searching or clustering are based on matching words as collections of digits devoid of semantic content. This is most clearly seen in the Soundex techniques for word matching by truncation, described in Chapter 15, since the resulting four-character strings are clearly meaningless in themselves. Some work is in progress to develop IR systems which will take account of semantic content, and results seem promising, in that improved recall and relevance appears to be possible. Chapters 6 and 7 discuss the problems involved in doing this intellectually, but it may be that in the future computer systems may be able to simulate this approach.[29]

Expert systems

Users who come to an IR system with an enquiry lack information that they need, but may not be able to express their need clearly: if they knew the question, they would be well on the way to finding the answer. It is during the reference interview that an intermediary tries to elicit information from the enquirer which will clarify the enquiry. At the other end of the process, the information which satisfies the enquiry will probably have come from one or more experts, who are knowledgeable in the subject. We may be able to help enquirers by developing computer *expert systems*, in which we store information gathered from experts together with rules and procedures to enable the users to get to the information they need despite starting from a position of ignorance.[30] The expert system is thus intended to parallel the purpose of the reference interview, but also to eliminate the steps of reference retrieval and document retrieval by providing direct answers.

In constructing an expert system for a particular subject we face certain problems. The first of these is that of gathering available information within the carefully defined scope of the system. This may be begun by a literature search, which is likely to identify those who may be regarded as experts in the field. The next step is to consult the experts themselves, and it here that we meet the second problem. The experts should be able to confirm the accuracy and adequacy of the information we have gathered, but they may find it very much more difficult to explain how they themselves acquired the information. Over the course of years we all develop mental information gathering and processing habits which enable us, when faced with a problem, to come up with a solution heuristically; we make decisions based on past experience without identifying each step of the thought processes which lead us to the answer. For a computer program to function, each step must be clearly identified and set down, otherwise the program will not be able to perform the task for which it is intended. A third problem is that in finding answers to questions we do not rely solely on knowledge specific to the subject, but use a wide range of general knowledge to provide us with context, analogies and instances which help us to make decisions which enable us to reach our goal. It is not practical to incorporate the whole of this range of general knowledge into an expert system; in order to make the process manageable we have to limit the information we put into the system to that which is specific to the subject area covered. Enquirers, on the other hand, may well stray outside these narrow bounds, if only because they are not aware of them.

Once we have established the knowledge base for a system, and elucidated the decision-making processes used by experts, we still have to incorporate what we have learnt into a computer system, using one of the programs already written for expert system development, and design a suitable user interface, bearing in mind the target group for whom the system is intended. Once a prototype has been constructed it has to be tested and, almost certainly, modified to correct any imperfections. Ardis gives an example of a difficulty arising in the design of an expert system to help users in online patent searching; one of the problems which was not recognized at the planning stage was that many users did not appreciate the difference between a patent and a trademark. The reference librarians who normally answered these enquiries would of course have implicitly recognized the two kinds of enquiry as separate, and this had to be built into the prototype once it had been made explicit by failures with the system. The features of and requirements for an expert system are summarized as follows:

The expert system:

1 must represent the expert's domain-specific knowledge in the way that the expert uses that knowledge
2 must incorporate explanation processes and ways of handling uncertainty
3 typically pertains to problems that can be symbolically represented
4 is more tolerant of user errors than conventional programs

In order to achieve this:

1 there must be at least one acknowledged expert in the subject area
2 the sources of the expert's expertise are judgement and experience
3 the expert must be able and willing to explain his/her knowledge
4 the problem must be well-bounded
5 the problem area must have a real consensus
6 test data must be easily available.[31]

So far, only a few working systems exist in library and information science. There are systems to help map cataloguers, for reference work, and for evaluating donations, but most are very specialized. One related to another subject area is PLEXUS, which is intended for public library clients wanting gardening information.[32] We have yet to see any substantial transfer of reference work from people to computers, but this will no doubt be a future trend, now that the proponents of artificial intelligence have accepted the present limitations on their work and are concentrating on what can be achieved.

Summary

This chapter has attempted to give an overview of the use of online searching, some of the background to its present role, and an indication of the kind of techniques that can be used. The number and scope of online databases means that online searching is now the normal way of finding information for many people. By no means all databases are bibliographic; there are financial databases, for example, to enable us to gamble on the stock market, if we wish, from the comfort of our own home! Increasingly, databases are including information other than text; as mentioned earlier, technology now allows us to retrieve graphics and sound. Chemical databases have included structural diagrams for compounds for many years, but the graphics involved are very simple compared with what is now available. Statistical databases are widely used; census data is becoming available to industry and commerce, as well as to the general public, much faster now that it is computer-compiled. The first edition of the *Oxford English dictionary* took 40 years to compile, in 13 volumes, and a supplement published five years later became necessary as a result of the extended editing process; the second edition is now available on one CD-ROM, and took six years to produce. Children are becoming accustomed to using computers and CD-ROM sources at school and at home, and will expect to find the same kind of information available for work purposes later in life. Yet we have seen that controlled vocabularies compiled by intellectual effort are still frequently used to achieve satisfactory results. One small experiment showed that natural language gave higher precision but lower recall than the use of a controlled vocabulary.[33] Making both available gave the user the option of high recall or high precision. We look at some of these controlled vocabularies in the following chapters on assigned indexing.

References

1 *Classification research for knowledge representation and organization: proceedings of the 5th International study conference on classification research, Toronto, Canada, June 24–28 1991*, Williamson, N. J. and Hudon, M. (eds.), Elsevier, 1992. (FID 698)

2 Austin, C. J., *MEDLARS, 1963–1967*, Bethesda, MD, National Library of Medicine, 1968. The MEDLINE database now contains several million references.

3 *Gale directory of databases*, Detroit, Gale Research Inc, 1995. In this edition, v1 lists over 5300 online databases; v2 lists 2015 CD-ROM products and another 2200 databases available on floppy disk, magnetic tape and other media.

Tenopir, C., 'Full-text databases', *Annual review of information science and technology,* **19**, 1984, 215–46.

4 Larson, S. E. and Williams, M. E., 'Computer assisted legal research', *Annual review of information science and technology,* **15**, 1980, 251–86.

5 Summit, R. K., 'In search of the elusive end user', *Online review,* **13** (6), 1989, 485–91.

6 Cornick, D., 'Being an end user is not for everyone', *Online,* **13**, March, 1989, 49–54.

Fisher, J. and Bjorner, S., 'Enabling online end-user searching: an expanding role for librarians', *Special libraries,* **85** (4), Fall 1994, 281–91.

Harman, D., 'User-friendly systems instead of user-friendly front-ends', *Journal of the American Society for Information Science,* **43** (2), 1992, 164–74. Suggests that implementing user-friendly front-ends is an inadequate substitute for improving the power of search engines.

7 Hartley, R. J., Keen, E. M., Large, J. A. and Tedd, L.A., *Online searching: principles and practice*, London, Bowker Saur, 1990.

8 Dalrymple, P. W. and Roderer, N. K., 'Database access systems', *Annual review of information science and technology,* **29**, 1994, 137–78.

9 ANSI Z39.58:1992 *Common command language for online information retrieval*, Bethesda, MD, National Information Standards Organization, 1992.

10 Armstrong, C. J. and Large, J. A. (eds.), *Manual of online search strategies*, Boston, Mass., G. K. Hall, 1988.

11 'Public access online catalogs', Markey, K. (ed.), *Library trends,* **33** (4), 1987,

523–67. (The point is made here in relation to OPAC searching, but it is of course generally valid.)

12 Keen, E. M., ‘The use of term position devices in ranked output experiments’, *Journal of documentation,* **47** (1), 1991, 1–22.

Keen, E. M., ‘Some aspects of proximity searching in text retrieval systems’, *Journal of information science,* **18** (2), 1992, 89–98.

13 Walker, S., ‘Evaluating and enhancing an experimental online catalogue’, *Library trends,* **35** (4), 1987, 631–45.

14 Sievert, M. and Boyce, B. R., ‘Hedge trimming and the resurrection of the controlled vocabulary in online searching’, *Online review,* **7** (6), 1983, 484–94.

15 Fidel, R., ‘Thesaurus requirements for an intermediary expert system’, *in Classification research for knowledge representation and organization: proceedings of the 5th International study conference on classification research, Toronto, Canada, June 24–28 1991*, Williamson, N. J. and Hudon, M. (eds.), Elsevier, 1992, (FID 698), 209–13.

16 Blair, D. C., *Language and representation in information retrieval,* New York, NY, Elsevier Science Publishers, 1990.

17 Cleverdon, C. W. ‘Optimizing convenient online access to bibliographic databases’, *Information services and use*, **4** (1–2), 1984, 37–47.

Cleverdon, C. W. [letter to the editor] *Online review,* **14**, 1990, 35, suggests that intermediaries support Boolean searching because it needs them to make it practical!

Pape, D. L. and Jones, R. L., ‘STATUS with IQ: escaping from the Boolean straitjacket’, *Program,* **22** (1), 1988, 32–43.

18 Maron, M. E. And Kuhns, J. L., ‘On relevance, probabilistic indexing and information retrieval’, *Journal of the Association for Computing Machinery,* **7** (3), 1960, 216–44.

Maron, M. E., ‘On indexing, retrieval and the meaning of about’, *Journal of the American Society for Information Science,* **28** (1), 1977, 38–43.

19 Doszkocs, T. E., ‘An associative interactive dictionary (AID) for online bibliographic searching’, in *The information age in perspective: proceedings of the ASIS annual meeting, November 1978.* White Plains, NY, Knowledge Industry Publications, 1978, 105–9.

20 Svenonius, E., ‘Classification: prospects, problems and possibilities’, in *International study conference on classification research, Toronto, Canada, June 24–28 1991,* Williamson, N. J. and Hudon, M. (eds.), Elsevier, 1992. (FID 698), 5–25.

21 Salton, G. (ed.), *The SMART retrieval system: experiments in automatic document processing*, Englewood Cliffs, NJ, Prentice-Hall, 1971.

Salton, G. and McGill, M. J., *Introduction to modern information retrieval,* New York, NY, McGraw-Hill, c1983, Chapter 3.

Salton, G. and Buckley, C., ‘Improving retrieval performance by relevance feedback’, *Journal of the American Society for Information Science,* **41** (4), 1990, 288–97.

Kantor, P. B., ‘Information retrieval techniques’, *Annual review of information*

science and technology **29**, 1994, 53–90.

22 Doszkocs, T. E., and Rapp, B. A., 'Searching MEDLINE in English: a prototype user interface with natural language query, ranked output, and relevance feedback' in *Information choices and policies, proceedings of the ASIS annual meeting, 1979*, White Plains, NY, Knowledge Industry Publications, 1980, 131–9.

23 Maron, M. E. and Kuhns, J. L., 'On relevance, probabilistic indexing and information retrieval' *Journal of the ACM,* **7** (3), 1960, 216–44.
Maron, M. E. 'On indexing, retrieval and the meaning of about', *Journal of the American Society for Information Science,* 28 (1), 1977, 38–43.

24 Lancaster, F. W., *Evaluation of the MEDLARS demand search service*, Bethesda, MD, National Library of Medicine, 1968.

25 Vickery, B. C. and Vickery, A., 'Online search interface design', *Journal of documentation,* **49** (2), 1993, 103–87.

26 Van Rijsbergen, C. J., *Information retrieval*, 2nd edn, London, Butterworths, 1979.

27 Sparck Jones, K., *Automatic keyword classification for information retrieval*, London, Butterworths, 1971.
Needham, R. M. and Sparck Jones, K., 'Keywords and clumps: recent work on information retrieval at the Cambridge Language Research Unit', *Journal of documentation,* **20** (1), 1964, 5–15. Included in *Theory of subject analysis . . .*

28 Salton, G. and McGill, M. J., *Introduction to modern information retrieval*, New York, NY, McGraw-Hill, c1983, Chapter 6, section 4.

29 Sembok, M. T. and van Rijsbergen, C. J., 'SILOL: a simple logical-linguistic document retrieval system', *Information processing & management,* **26** (1), 1990, 111–34.

30 Poulter, A., Morris, A. and Dow, J., 'LIS professionals as knowledge engineers', *Annual review of information science and technology,* **29**, 1994, 305–50.
Vickery, B. C., 'Knowledge representation: a brief review', *Journal of documentation',* **42** (3), September 1986, 145–59.
Alberico, R. and Micco, M., *Expert systems for reference and information retrieval*, Westport, CT, Meckler, 1990.
Artificial intelligence and expert systems: will they change the library?, Lancaster, F. W. and Smith. L. C. (eds.), Urbana-Champaign, University of Illinois Graduate School of Library and Information Management, 1992. (Clinic on library applications of data processing: 1990)

31 Ardis, S. B., 'Online patent searching: guided by an expert system', *Online,* **14** (2), March 1990, 56–62.

32 Vickery, A. *et al.,* 'A reference and referral system using expert system technique', *Journal of documentation,* **43** (1), March 1987, 1–23.

33 Rowley, J. E., 'A comparison between free language and controlled language indexing and searching', *Information services and use,* **10** (3), 1990, 147–55.

Chapter 6

Assigned indexing 1: Semantics

In Chapter 3 we looked at ways in which printed indexes could be derived from information manifest in a document. In Chapter 5, we considered some of the ways in which files may be searched online, again using the information manifest in the document, e.g. titles, abstracts or full text; the discussion indicated some of the problems that are likely to arise in doing this, and we referred in passing to the use of 'controlled vocabularies' to assist in solving these problems, without at that time showing what was meant by a controlled vocabulary. We have also seen in Chapter 2 that full-text searching gives the highest possible level of exhaustivity, which tends to be associated with high recall but low relevance; we may wish to have some method of summarization to supplement the depth indexing of text searching. A discussion of these problems leads to the idea of *assigned indexing*.

Firstly, we have to choose the words which we will use in a search of the system by trying to think of all the words that the authors of the documents we have indexed might have used to describe the topic we are interested in, and, having chosen the words, we have to think of the various forms in which they might occur. Truncation serves as a means of merging different word forms, but not always; TEACH* will retrieve teaching and teacher but not taught. Secondly, we often need to search for combinations of terms; word pairs are more significant than the individual words on their own, but we often find ourselves wanting to associate more than two words. This process of *coordination* is, as we have seen, a process of class intersection. To use one of our previous examples, our collection of documents (the universe of discourse) contains a set of documents containing the word 'water', and each of these sets forms a class, and if we are searching for documents on 'water pollution' we are looking for the intersection of these two classes. Any process involving class intersection is likely to be a powerful method of reducing the total number of documents retrieved; we have also seen that a reduction in recall is often accompanied by an improvement in relevance, so we would expect *coordination* to be a useful method of obtaining improved relevance. On the other hand, class union ('A' OR 'B') increases the total number of documents retrieved, so we would expect the inclusion of alternative terms to be a device for improving recall.

We also noted that *water* on its own might not retrieve all the documents of interest, because they might use different but related terms: *sea* and *river*. We noted in Chapter 5 that computer matching does not involve semantic content, so that it cannot lead to related terms directly; even computer classification is based on such factors as co-occurrence, not similarity of meaning. In order to carry out an adequate search of our collection of documents, we had to think of not only the words in which we were interested, and all the forms in which they might be used, but also all the alternative or related forms. We then had to decide just how we were going to coordinate these words in order to retrieve relevant documents, while at the same time excluding words or combinations of words which would retrieve irrelevant material. This is obviously quite a complex operation, and if we are to do it well we need some guidance: a list of words showing their relationships and indicating ways in which they might usefully be combined to give the class intersections we are interested in. However, in Chapter 1 we pointed out that what we are actually trying to do is carry out a matching operation between the messages which in their encoded form are the input to our system and the messages – also in their encoded form – which represent the questions we put to the system. This concept of matching is of course strongly reinforced by our examination of computer-based systems, which depend on the computer to match the words of our question against the words in the documents.

Now if we are to use a list of words to help us in our searching, it would appear that we would increase the chances of achieving successful matches if we used the same list of words to encode the document at the input stage, and *assigned* the appropriate words to the documents ourselves rather than rely on the authors' choice. In other words, we devise an *indexing language* and use this for both encoding operations: input and question. Such systems are referred to as *assigned indexing* systems, and most of the rest of this book is devoted to the problems of constructing and using such systems. In this chapter we examine some of the basic theoretical problems.[1]

Choice of terms

Assigned indexing is also known as *concept indexing*, because what we are trying to do is to identify the concepts involved in each document. (Concept: idea of a class of objects; general notion.)[2]

One analysis suggests that there are five categories of concept: entities; activities; abstracts; properties; heterogeneous. A concept is denoted by a *term* which may consist of more than one word. (Term: a word or expression that has a precise meaning in some uses, or is peculiar to a science, art, profession or subject.)[3] We may examine each of these categories in more detail. *Entities* are things which may be given a denotative meaning,[4] i.e. we can identify them by pointing at them. They may be physical, e.g. matter, or physical phenomena; chemical, e.g. molecular states, minerals; biological, e.g. living being; or artefacts, i.e. manufactured items. *Activities* are usually denoted by verbal nouns, e.g. building, lubricating, though in some cases we find the passive rather than the active form, e.g. lubrication. *Abstracts* usually refer to qualities or states, and are given connotative meanings, i.e. each of us may attribute a different meaning to them depending on our particular corpus of experience. They may be physical, e.g. energy; symbolic, e.g. Justice as a blindfolded figure; or behavioural, e.g. truth (the definition of which was questioned on at least one notable occasion). *Properties* are of two kinds, which are distinguished by their grammatical form. Adjectival forms can only be used in conjunction with a noun, which they qualify in a subjective or attributive way, once again giving a connotative meaning. They may relate to sight, e.g. dull, shiny, symmetrical; sound, e.g. loud, musical; or to the other three senses, touch, taste and smell. They may also relate to mechanical properties, e.g. loose, rigid. Noun forms describe physical properties which may be measured, e.g. rigidity, reflectivity, loudness. It will be clear at once that there will be in many cases a definite relationship between the two kinds. We may refer to the rigidity of an iron bar, for example, in which case we are thinking of the property; or we may refer to a rigid bar, in which case we are using the property to define the kind of entity that we are considering.

Heterogeneous concepts form a very mixed bag, in that they usually represent concepts which might be further analysed into two or more simpler concepts which would fit into the other categories, but are nevertheless regarded as unitary concepts and treated as such. Willetts[5] has suggested some types:

Roles of man (Entity + Activity, Entity + Property) e.g. teacher, landlord
Groups of man (Entity + Abstract) e.g. society, conference
Types of building (Entity + Activity + Property) e.g. library, theatre
Discipline (all four) e.g. Physics, Medicine
Groups of chemicals (Entity + Activity, Entity + Property) e.g. catalysts, polymers

Austin[6] would regard the Groups as Aggregates, while the rest would fit into most indexing systems quite smoothly as they stand. Are there any advantages to further analysis?

During the 1950s a team at Case Western Reserve University worked on a system of analysis known as semantic factoring.[7] The objective was to break down every concept into a set of fundamental concepts called semantic factors. Because of their fundamental nature, there would only be a limited number of these factors. A concept would be denoted by the appropriate combination of semantic factors, and the use of a complex set of roles and links enabled the indexer to write a ‘telegraphic abstract’ which would represent the subject of a document in a computer file.

The method is clearly a powerful one, but is open to some doubts and objections. Exactly how far does one carry such an analysis? Heat and temperature, for example, could be specified as *movement* of *molecules*. Again, it is possible to specify a concept by using only some of its attributes; or perhaps more significantly, is it ever possible to specify *all* the attributes for a given concept? For example, thermometer may be specified as instrument: measuring: temperature, and barometer may be specified as instrument: measuring: pressure. Neither reveals the fact that both may have other factors in common, for example the fact that they may be mercury-in-glass devices. Certainly for most purposes a mercury barometer has more in common with an aneroid barometer than it does with a thermometer, but this may not be the case if we are thinking of the instrument maker. If we start to think of a particular individual, we may have no difficulty in putting a name to the object of our thoughts; we may find it impossible in practical terms to think of all the possible terms that might be needed to specify an individual without naming them. Sex, age, nationality, family status, marital status, height, weight, occupation, language, religion – the list is almost endless. Furthermore, we may find ourselves in the position of not knowing all of the information we require; we have to remember that we are dealing with the information in a collection of documents, and this will usually be incomplete.

We also have a problem in analysing certain concepts which lose their significance if split up into their constituent parts. A soap opera is not a kind of opera, nor is it a form of soap;[8] 'moment of truth' cannot be analysed further; a blackbird is a specific species of bird, but there are many black birds;[9] a rubber duck is not a species of duck (despite evidence presented in some TV advertising) and nowadays it is rarely made of rubber.[10] Fortunately, though the theoretical problems involved may not all have been solved, in practice, solutions which are reasonably effective *can* be found.

Choice of form of word

During the above discussion of categories of concept it should have become apparent that – with the sole exception of adjectival properties, which cannot stand alone – all the concepts involved were denoted by nouns. Even activities are denoted by verbal nouns, active or passive, e.g. cataloguing and classification. In fact, it is the norm in indexing languages to use nouns as far as possible, and various sets of rules have been drawn up to give guidance on the use of singular and plural. Table 6.1 is based on the rules given in the EJC *Thesaurus of engineering and scientific terms*, described in Chapter 25, while the ISO,[11] BSI[12] and ANSI[13] have all published standards on thesaurus construction. A useful rule of thumb is: how much? – use the singular; how many? – use the plural.

Table 6.1 Choice of singular or plural form of noun

Type of term	Use singular	Use plural
Materials Properties	When specific, e.g. polythene density	When generic, e.g plastics chemical properties
Objects Events Objects specified by purpose		Cars Laws Wars Lubricants
Processes Proper names Disciplines Subject areas	Lubricating Earth (the planet) Law War	

Homographs

The same spelling is sometimes used for different words, which may or may not be pronounced the same, e.g. sow and sow, China and china. This may arise from a figure of speech such as metonymy or synecdoche, in which we use part of a description to mean the whole; it may be through analogy, when terms such as 'filter' from hydraulic engineering are used by electrical engineers; or it may simply be an etymological accident. Whatever the cause, there is likely to be confusion if we do nothing to distinguish such words. One way of doing this is to qualify each by another word in parentheses to show the context and thus the meaning, e.g.:

PITCH (Bitumen)
PITCH (Football)
PITCH (Music)
PITCH (Slope)

If we do not distinguish homographs we shall get reduced relevance; the seriousness of this will depend on the coverage of our system. For example, if our system only covers music there will be no problem with the word pitch, since other meanings than the musical are unlikely to arise at the input stage. However, it has been pointed out[14] that the 20 most frequently used English nouns have an average of seven meanings each, so we must obviously be aware of the problem.

Relationships

We have seen that in addition to the choice of terms and the form in which they should be used, there are two kinds of relationship between terms that we have to take into account: the recognition of terms denoting related subjects such as water, sea and river, and the association of otherwise unrelated terms to represent composite subjects. One place where we can identify the kinds of terms used by authors, and how they are associated is in the titles they give their works. If we study Table 6.2 carefully, we can see first of all that the titles, which are taken from the Library of Congress catalogue, fall into three major groups: Education, Agriculture and Cookbooks. Within each of these three subject areas we can see examples of both kinds of relationship, and it can be seen that one kind is permanent, and arises from the definitions of the subjects involved, while the second kind arises from the associations we find in documents, and represent temporary, *ad hoc*, associations. The first kind are known as semantic relationships: corn is always a kind of cereal. The second are called syntactic: Disinfestation is an activity carried out on a crop, in this instance grain, an entity; coloured immigrants are people, entities, being educated, activity, in Britain, place. This suggests that our indexing language must contain the equivalent of a dictionary, to show semantic relationships, and a grammar, to cater for syntactic relationships. In computer searching, the grammar may be the rudimentary provision of Boolean logic, but in printed indexes or shelf arrangement we may wish to show more complex relationships.

Semantic relationships

We find that these may be considered in three groups: equivalence, hierarchical and affinitive/associative. The first two groups are reasonably straightforward, but the third is much less clearly defined, and is the group which causes most problems in practice.

We can examine the various kinds of each of the three groups in more detail, with most of our examples taken from the titles in Table 6.2.

Table 6.2 Related subjects

Concept analysis	
1	Education of women in India 1921–1966.
2	Acceleration and the gifted.
3	The costs of education.
4	The teaching of Physics at university level.
5	Teaching French: an introduction to applied linguistics.
6	Saga of the steam plow. [plough]
7	The main course cookbook.
8	The corn earworm in sweet corn: how to control it.
9	Wheat.
10	The potato.
11	New first year mathematics: teacher's book.
12	Radiation disinfestation of grain.
13	The education of coloured immigrants in Britain.
14	Modern corn production.
15	The elementary school: a perspective.
16	Agricultural financing in India.
17	Technology of cereals.
18	Meat, fish, poultry and cheese . . .
19	A cyclopedia of education. [i.e. encyclopedia]
20	Soups and hors d'oeuvres.
21	Curriculum theory.
22	A world of nut recipes from soups to savories.
23	Economic aspects of higher education.
24	The pecan cookbook.
25	Education improvement for the disadvantaged in an elementary setting.
26	The evolution of the comprehensive school.
27	The world book of pork dishes.
28	New media and college teaching.
29	Potatoes in popular ways.
30	Educational aids in the infant school
31	The planetarium: an elementary School teaching resource.
32	Vegetable cookbook.
33	English in the primary school.
34	Talking about puddings.

Equivalence

Synonyms and antonyms
Quasi-synonyms
  Same continuum
  Overlapping
Preferred spelling
Acronyms, abbreviations
Current and established terms
Translations

The English language is rich in synonyms and near-synonyms, because it has roots in both Teutonic and Romance languages. While it is true that Wordsworth's ode would sound less impressive as *Hints of deathlessness* than as *Intimations of immortality*, the former is as correct a statement of the subject as the latter. Many subjects have both a common name and a scientific name: potato and *Solanum tuberosum*; American usage differs from British or Australian: elementary school and primary school; authors differ in their usage: the word college is used in more than one sense. By not merging synonyms, we shall be separating literature for the lay reader from that for the expert, American from British, one author from another; this is likely to improve relevance at the expense of recall. By merging synonyms, we are likely to improve recall at the expense of relevance.

It may seem odd to include antonyms with synonyms, yet in trying to retrieve information we may often find it useful to treat them in the same way. In practice, gifted children are often disadvantaged! Quasi-synonyms often represent points on the same continuum, or the overlapping of concepts. Antonyms and quasi-synonyms may overlap; roughness and smoothness may be thought of as antonyms, but they lie on a continuum which often represents a subjective judgement. We may find difficulty in clearly distinguishing pre-school, infant school and primary school; comprehensive schools overlap secondary schools; economics, costs and financing are often not clearly distinguished; the distance between two points is a length.

The other four instances are self-explanatory: plow (US) = plough (UK), labor (US) = labour (UK); ERIC = Educational Resources Information Clearinghouse; Third World, Developing Countries, Underdeveloped Countries; Zhurnal = Journal.

The equivalence relationship implies that there will be more than one term denoting the same concept. In a controlled vocabulary it is usual to select one term as the *preferred term*, and use only that one in our indexing. We must of course make provision for those users who look for information under one of the other terms, and this is discussed below in the section on showing semantic relationships.

Hierarchical

Genus – species
Whole – part

The usual kind of hierarchical relationship is that of genus to species, which represents class inclusion (all A is B; some B is A). It is seen most clearly in the biological sciences (all mammals are vertebrates; some vertebrates are mammals), but is also found in other subject fields; indeed, much of classification is concerned with the establishment of hierarchies. Austin[15] distinguishes what he calls quasi-generic relationships from true generic, using the criterion of permanence; a potato is always a plant of the species *Solanum tuberosum*, but it may appear on our dinner plate as part of a meal, or it may be used by children to print out simple designs. A planetarium is a kind of teaching resource. Beef, veal and pork are kinds of meat.

Whole-part relationships are not generic. A wheel is not a species of bicycle, nor is a door a species of house. However, it is convenient to regard whole-part relationships as hierarchical, and it has been recommended that the two types should be distinguished as Genetic and Partitive.[12] The partitive relationship is illustrated by four particular examples:

a systems and organs of the body
b geographic locations
c disciplines or fields of discourse
d hierarchical social structures.

In each case, the name of the part should imply the name of the whole regardless of context, so that the terms can be organized as logical hierarchies.

Affinitive/Associative relationships

Coordination
Genetic
Concurrent
Cause and effect
Instruments
Materials
Similarity

Because these are the least well-defined, and often are not immediately obvious, they are the group most likely to cause problems in an indexing language. Indeed, Coates[16] criticised LCSH for including these relationships in what appears to have been a quite haphazard way. (More recently, new relationships within LCSH have been restricted to equivalence and hierarchical types.) Despite the difficulties, we should make some attempt to cater for these relationships by first of all recognizing that they exist and then trying to identify them systematically.

Some present fewer problems than others. Coordination is in effect a by-product of the generic relationship: species of the same genus are coordinate. Thus wheat and corn are both kinds of cereal crop; savo[u]ries, hors d'oeuvres, soups, entrees, main courses and puddings may be regarded as sequential courses of a meal. It is worth noting that if division of this kind is *dichotomous*, i.e. into A and A′, the result is to give two concepts which are antonyms, e.g. male and female, poetry and prose. For this reason antonyms are sometimes considered to fall into the associative rather than the equivalence group.

Genetic relationships are also straightforward, e.g. mother–son; here again we

Table 6.3 Relationships and associations

Relationships discussed	Word associations
Word forms	Word derivatives
synonyms	similar
antonyms	contrast
hierarchical	superordinate subordinate
coordinate	coordinate
whole-part	whole-part
cause and effect	cause and effect
instruments	verb-object
materials	material
similarity	similarity
genetic	—
—	assonance

may note that the first level of genetic division will give coordinate concepts, e.g. son–daughter. Concurrent refers to two activities taking place at the same time in association, and is thus open to much broader interpretation; an example is education–teaching. Cause and effect are rather easier to identify, though of course they have been the subject of much philosophical discourse; an optimistic example from the draft British Standard was teaching–learning. This has been replaced in the final version by the more prosaic but rather more solid diseases–pathogens. Instruments, e.g. teaching–media; and materials, e.g. plastic film–transparencies, are usually fairly obvious. The final category, similarity, is perhaps the most difficult of the affinitive relationships in that it necessarily implies a subjective judgement; how similar do two concepts have to be for us to recognize the relationship? We should not expect any great degree of consistency between different indexing languages.

It is interesting to compare the kinds of relationship discussed here with a similar categorization of relationships revealed by psychological word associations.[17] It may be that a study of such associations may throw further light on the kinds of relationship that we need to cater for in our index vocabularies (Table 6.3)

The need to recognize semantic relationships

At the beginning of this chapter we saw how the need to identify semantic relationships might arise, but in view of the fairly detailed analysis that we have just carried out, it is worth restating the problem, from two rather different points of view. We started off from the viewpoint of the searcher trying to carry out a search in a computer-based system using the texts, or parts of the texts, of the documents in our collection. Just which collection of terms do we have to use to ensure that we have covered all the possible approaches to a concept? To put it another way: if the term we first think of does not retrieve the documents we want (or perhaps does not retrieve any documents at all!) *what other terms can we substitute*? It is obviously