1

Basic concepts of information retrieval systems

Introduction

The term 'information retrieval' was coined in 1952 and gained popularity in the research community from 1961 onwards.[1] At that time the organizing function of information retrieval was seen as a major advance in libraries that were no longer just storehouses of books, but also places where the information they hold is catalogued and indexed.[2] Subsequently, with the introduction of computers in information handling, there appeared a number of databases containing bibliographic details of documents, often married with abstracts, keywords, and so on, and consequently the concept of information retrieval came to mean the retrieval of bibliographic information from stored document databases.

Information retrieval is concerned with all the activities related to the organization of, processing of, and access to, information of all forms and formats. An information retrieval system allows people to communicate with an information system or service in order to find information – text, graphic images, sound recordings or video that meet their specific needs.

Thus the objective of an information retrieval system is to enable users to find relevant information from an organized collection of documents. In fact, most information retrieval systems are, truly speaking, document retrieval systems, since they are designed to retrieve information about the existence (or non-existence) of documents relevant to a user query. Lancaster[3] comments that an information retrieval system does not inform (change the knowledge of) the user on the subject of their enquiry; it merely informs them of the existence (or non-existence) and whereabouts of documents relating to their request. However, this notion of information retrieval has changed since the availability of full text documents in bibliographic databases. Modern information retrieval systems can either retrieve bibliographic items, or the exact text that matches a user's search criteria from a stored database of full texts of documents. Although information retrieval systems originally meant text retrieval systems, since they were dealing with textual documents, many modern information retrieval systems deal with multimedia information comprising text, audio, images and video. While many features of conventional text retrieval systems are equally applicable to multimedia information retrieval, the specific nature of audio, image and video information has called for the development of many new tools and techniques for information

retrieval. Modern information retrieval deals with storage, organization and access to text, as well as multimedia information resources.

Features of an information retrieval system

Figure 1.1 presents the conceptual view of an information retrieval system. An information retrieval system is designed to enable users to find relevant information from a stored and organized collection of documents. Thus the concept of information retrieval presupposes that there are some documents or records containing information that have been organized in an order suitable for easy retrieval. The documents or records we are concerned with contain bibliographic information, which is quite different from other kinds of information or data. We may take a simple example. If we have a database of information about an office or a supermarket, all we have are the different kinds of records and related facts, such as, for an office, names of employees, their positions, salary and so on; in the case of a supermarket, names of different items, prices, quantity and so forth. The retrieval system here is designed to search for and retrieve specific facts or data, such as the salary of a particular manager, or the price of a certain perfume. Conventional database management systems, such as Access, Oracle, MySQL, and so on, deal with structured data, where the organization or structuring of data takes place depending on the specific attributes of the data elements. For example, in a database of university students, the various data elements could be the attributes of specific student records, such as student registration number, student name, address, subjects studied, grades and so on. In contrast to this, a database of items sold in a supermarket could be the name of the item with its barcode, manufacturer, supplier, price and so forth. So, the first database in this example will be structured according to the specific attributes of students, while in the second case the database will be structured according to the attributes of specific products. The particular objective of these databases is to allow the user to search for specific records that match one or more specific conditions or search criteria, for example, details of a certain student with a particular registration number; details of a specific product with a particular barcode; a list of all the students that are registered for a specific course; or the products of a particular type within a certain price range, for example toothpaste that costs between one and four pounds.

As opposed to a conventional database management system, an information retrieval system is designed to deal with unstructured data. The major objective of an information retrieval system is to retrieve the information – either the actual information or the documents containing the information – that fully or partially match the user's query. The database may contain abstracts or full texts of documents, such as newspaper articles, handbooks, dictionaries, encyclopedias, legal documents, statistics and so on, as well as audio, images and video information. Whatever the nature of the database may be – bibliographic, full-text or multimedia – the system presupposes that there is a group of users for whom the

system is designed. Users are considered to have certain queries or information needs, and when they put forward their requirement to the system, the latter should be able to provide the necessary bibliographic references of those documents containing the required information; some systems also retrieve the actual text, image, table or chart relevant to the information needs of the user.

It will be easy to understand the basic functions of an information retrieval system if we take the following simple example. Let us imagine that we want to find information about a term, say 'internet', in a book. One approach would be to begin with the first word in the first sentence in the book, and continue to look for the term 'internet' until we find it or we reach the end of the book. However, in real life, we don't do this. Instead, we use an index – the 'back-of-the-book index' – to look for a match for the search term, and if we find a match then we take note of the corresponding references – the page number(s) where the term occurs – and we move to the specific page(s) to find the information. In their simplest form, most information retrieval systems work in this way.

Although historically information retrieval systems were designed to help people find information from bibliographic and textual databases, in today's world we use information retrieval systems in almost every aspect of our daily lives, for example, to retrieve a message or e-mail received or sent on a specific date; to find messages sent to or by a particular person; to find something or someone on the web; to search for a book in an online library catalogue or in a digital library; to search for a song or to find a video on YouTube; and so on. The following are some typical activities where we use information retrieval systems, in some form or other, in our day-to-day life and activities:

- to search for information resources in a library's online public access catalogue (OPAC), which provides access to the library's collections
- to search for information in online bibliographic or full-text databases (database search services) such as Dialog (www.dialog.com), Ovid (www.ovid.com) or ABI/Inform (www.proquest.com/products_pq/descriptions/abi_inform.shtml), providing access to remote collections
- to access e-books and e-journal services such as NetLibrary (www.netlibrary.com/), Emerald (www.emeraldinsight.com), and Ingenta (www.ingenta.com), providing access to electronic books and journal articles
- to search for an e-mail address, a specific message, a phone number or an address on a mobile phone or in e-mail services such as Outlook Express, Gmail, or Eudora
- to search for information on institutional intranets and databases, such as those created by companies and institutions providing access to various information resources created within the institution
- to access information on websites either by going directly to the web page, by entering the web address or Uniform Resource Locator (URL) of the site, or by using tools such as search engines like Google (www.google.com); meta search engines, which provide information from more than one search engine, such as

Dogpile (www.dogpile.com) and Mamma (www.mamma.com); specialty search engines that use special techniques for search and/or display of results, such as Clusty http://clusty.com) and Answers.com (www.answers.com); and directories such as Yahoo! (www.Yahoo.com)

- to access information on the web using subject gateways that provide access to selected web resources in one or more specific discipline(s), such as Intute: social sciences (www.intute.ac.uk/socialsciences), Intute: humanities (www.intute.ac.uk/humanities) and Intute: medicine including dentistry (www.intute.ac.uk/medicine)
- to access information in digital libraries, such as the American Computing Machinery (ACM) digital library (http://portal.acm.org/dl.cfm), the New Zealand Digital Library (NZDL; www.nzdl.org) and the Networked Digital Library of Theses and Dissertations (NDLTD; www.ndltd.org)
- to search for music on iTunes
- to search for information on social networking sites such as Facebook, Twitter and YouTube.

Elements of an information retrieval system

Figure 1.1 shows that an information retrieval system may comprise one or more different types of documents and can contain text as well as multimedia information. All the documents are processed to create an index, which is searched for retrieval of information. In its most simple form, this index can be considered as a back-of-

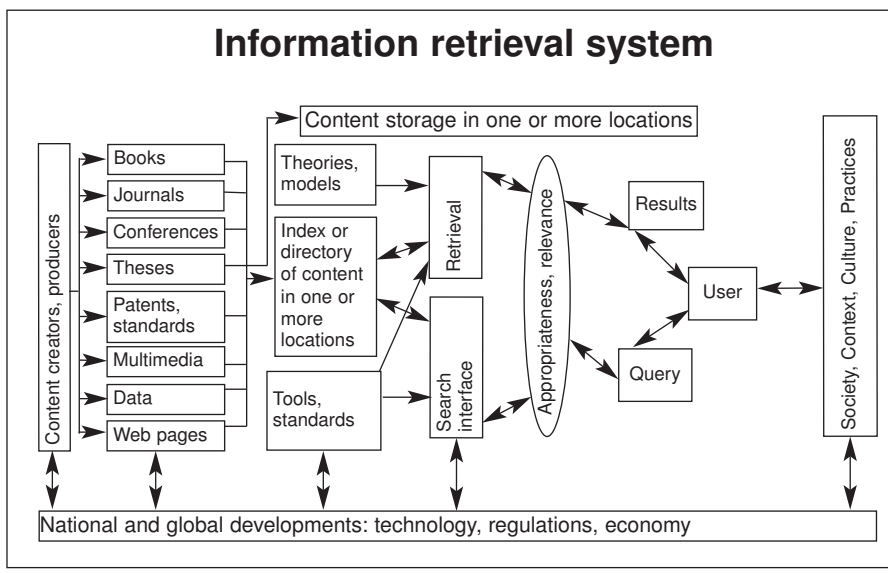

Figure 1.1 *Broad outline of an IRS*

the-book index, but in reality it is much more complex than that. Details of such indexes appear in Chapter 6. Some information retrieval systems, especially some web search tools (discussed in Chapter 18), use a directory, which is like a hierarchical list of subjects used to map documents in a collection, and which require users to browse through the directory to identify a preferred term or concept and follow the links from there to access the mapped documents. Details and examples of some web directories appear in Chapter 18. However, as in books, the actual documents in an information retrieval system are kept separately from the index, and it is the index that is used for an information search. Figure 1.1 shows that it is a rather complex process to create an index, and various tools, techniques and standards are used for the purpose. In information retrieval systems the documents and index may be located in one or more places in order to facilitate fast access and easy maintenance of the document and the index databases.

Users interact with information retrieval systems through an interface where they are usually expected to express their information needs in the form of a query, which is presented to the search system through a search expression that may contain one or more search terms presented in the form of a natural language sentence, or in a constrained natural language where search terms are linked with various search operators (details of query language and search operators appear in Chapter 9).

Information is retrieved (usually in the form of documents that contain the required information) whenever the search terms match the index terms; several information retrieval models and theories have been developed over the past five or so decades which are used for matching and retrieval. Details of these models appear in Chapter 9.

As may be noted from Figure 1.1, information search and retrieval processes are very much influenced by the concepts of appropriateness and relevance. One of the major problems that information retrieval systems based on query formulation face is that often users cannot express their information needs in the form of queries, and cannot pass them to the search system through appropriate search statements. Therefore, however sophisticated an indexing and retrieval system may be, the overall performance of the information retrieval system may not be satisfactory to users because some users do not make appropriate search statements. Again, the success of an information retrieval system very much depends on the user's judgement of whether retrieved documents are relevant to their query (which may or may not be a true reflection of an information need in the first instance). Therefore, information needs vis-à-vis query formulation and relevance judgement are a major area of study and research in information retrieval.

The box on the extreme right in Figure 1.1 shows that users, their information needs, information search behaviour, relevance judgements and so on can be influenced by a number of factors that are outside the realm of a specific information retrieval system, but are nonetheless very important – society, context, culture and practices – which all have an effect on users and their information-seeking and retrieval activities. This is an important area of study and research, details of which appear in Chapters 10 and 11.

Overall, the entire information retrieval system – the technical as well as the social and human aspects – are influenced significantly by several external factors including national and global developments in technology, regulations and the economy, and indeed these influences can be clearly seen through the development of information retrieval over the past 50 or so years since the introduction of computers in information retrieval, and more so over the past one and a half decades since the advent of the internet and the web.

Purpose

An information retrieval system is designed to retrieve the documents or information required by the user community. It should make the right information available to the right user. Thus, an information retrieval system aims to collect and organize information in one or more subject areas in order to provide it to users as soon as they ask for it. Belkin[4] describes how information retrieval systems are used in the following way:

- A writer presents a set of ideas in a document using a set of concepts.
- Somewhere there are users who require the ideas but may not be able to identify them; in other words, some people lack the ideas put forward by the author in their work.
- Information retrieval systems match the writer's ideas expressed in the document with the users' requirements or demands for them.

Thus, an information retrieval system serves as a bridge between the world of creators or generators of information and the users of that information. Hence some researchers comment that information retrieval is a communication process.[5]

Functions

An information retrieval system deals with various sources of information on the one hand and users' requirements on the other. It must:

- analyse the contents of the sources of information as well as the users' queries, and then
- match these to retrieve those items that are relevant.

Information retrieval systems have the following functions:

- to identify the information (sources) relevant to the areas of interest of the target users' community; this is a challenging job especially in the web

environment where virtually everybody in the world can be the potential user of a web-based information retrieval system
- to analyse the contents of the sources (documents); this is becoming increasingly challenging as the size, volume and variety of information sources (documents) is increasing rapidly; web information retrieval is carried out automatically using specially designed programs called spiders (discussed in Chapter 18)
- to represent the contents of analysed sources in a way that matches users' queries; this is done by automatically creating one or more index files, and is becoming an increasingly complex task due to the volume and variety of content and increasing user demands
- to analyse users' queries and represent them in a form that will be suitable for matching the database; this is done in a number of ways, through the design of sophisticated search interfaces including those that can provide some help to users for selection of appropriate search terms by using dictionary and thesauri, automatic spell checkers, a predefined set of search statements and so forth
- to match the search statement with the stored database; a number of complex information retrieval models have been developed over the years that are used to determine the similarity of the query and stored documents
- to retrieve relevant information; a variety of tools and techniques are used to determine the relevance of retrieved items and their ranking
- to make continuous changes in all aspects of the system, keeping in mind the rapid developments in information and communication technologies (ICTs) relating to changing patterns of society, users and their information needs and expectations.

Components

It is evident from the above discussion that on the one side of an information retrieval system there are the documents or sources of information and on the other there are the users' queries. These two sides are linked through a series of tasks. Lancaster[6] mentions that an information retrieval system comprises six major subsystems:

- the document subsystem
- the indexing subsystem
- the vocabulary subsystem
- the searching subsystem
- the user-system interface
- the matching subsystem.

All the tasks mentioned in Figure 1.1 can be brought under two major groups – subject/content analysis, and search and retrieval. Subject or content analysis

includes the tasks related to the analysis, organization and storage of information. The process of search and retrieval includes the tasks of analysing users' queries, creation of a search formula, the actual searching and retrieval of information. The major emphasis of this book is laid on these two areas. Researchers in the information retrieval world are engaged in developing suitable methodologies for both sets of operations. Developments in the technological world, especially in computer and communication technologies, have provided an additional impetus to the development of information retrieval systems. Researchers who are working on the storage side of the information retrieval system are engaged in designing sophisticated methods for identification and representation of the various bibliographic elements essential for documents, automatic content analysis, text processing and so on. On the other hand, researchers working on the retrieval side are attempting to develop sophisticated searching techniques, user interfaces, and various techniques for producing output for local as well as remote users. The recent emergence of the internet, particularly the world wide web (discussed in Chapter 18), has made a significant impact on the information retrieval environment.

Kinds of information retrieval systems

Information retrieval systems can be categorized in a number of ways. For example, one can group them into two categories: in-house and online.

In-house information retrieval systems are set up by a particular library or information centre to serve mainly the users within the organization. One particular type of in-house database is the library catalogue. OPACs provide facilities for library users to carry out online catalogue searches and to then check the availability of the item required.

By online information retrieval systems we mean those that have been designed to provide access to a remote database(s) to a variety of users. These are mostly commercial services, and there are a number of vendors that handle them. With the development of optical storage technology another type of information retrieval system appeared on CD-ROM (compact-disc read-only memory). Information retrieval systems based on CD-ROM technology are mostly commercial services, though there have been some free and in-house developments too. Basic techniques for search and retrieval of information from the in-house or CD-ROM and online information retrieval systems are more or less the same, except that the online system is linked to users at a distance through the electronic communication network.

Another, and perhaps more appropriate, grouping could be made on the basis of the content, purpose and functions of information retrieval systems. In this approach four distinct types of information retrieval systems can be identified:

- OPACs
- online databases

- digital libraries and web-based information services
- web search engines.

OPACs demonstrate some typical and limited features of information retrieval systems, which are designed for keeping in view the nature of the documents that they handle as well as the user and their specific purpose for the information search. OPACs typically allow users to search by using typical bibliographic keys such as authors' names, titles of documents, subject descriptors and keywords, and at the end of a search they produce a list of documents, with some bibliographic information and a call number (an artificial number often comprising several alphanumeric characters and punctuations), the latter being the link between the retrieved documents and their physical location in the library. These information retrieval systems usually have less sophisticated search and retrieval features than other categories of information retrieval systems.

Online information retrieval systems appeared in the early era of computer applications in information retrieval, and over the past five decades these systems have gone through several improvements in their search and retrieval features. Examples of typical online information retrieval systems include those that are available through various search service providers – companies that provide access to remote text and other types of databases – such as Dialog, Ovid and Factiva. These information retrieval systems are designed to work on many large live databases comprising millions of documents, and they provide very attractive information search and retrieval features. A unique characteristic of these systems, as opposed to other online services, specifically the web search engines, is that they are fee- or subscription-based services, and they provide access to peer reviewed, quality, often scholarly, information sources.

Recent developments in information and communication technologies have widened the scope of online information retrieval systems. The internet and world wide web have made information available for use by anyone virtually anywhere who has access to the appropriate equipment. This has led to the development of several digital libraries and web-based information services that can be accessed remotely through a web interface. These information retrieval systems are different from the typical online information retrieval systems, described in the previous paragraph, in that often they are free and can be accessed by virtually anyone through the web.

The web has given birth to another category of information retrieval system, which is unique and quite different from the three other categories of information retrieval systems discussed above. These are the web search engines that are designed to provide access to vast amounts of web information resources. These information retrieval systems have some typical characteristics – they are robust, designed only to enable the users to find web resources; they do not guarantee access to the resources they retrieve but, perhaps most importantly, they are free at the point of use.

Design issues

A system can be defined as a set of interacting components, under human control, operating together to achieve an intended purpose. Thus a system carries out processing on inputs to produce required outputs; the agents of this processing are people and machines.[6, 7, 8, 9]

System design may be viewed as a series of choices from which the designer selects each element and tries to fit it with the proposed objective of the system. Therefore, if a system is designed carefully, the designer must be aware of the choices that are to be made, and he or she must consider the consequences of making any available choice. The life-cycle approach to system design suggests the following basic stages in the life of a system:[10]

1. An analysis has to be conducted in order to establish the requirements of a system, and to learn the various options available.
2. Next comes the design phase, which eventually gives rise to a specific system to match the requirements.
3. Next comes the implementation stage, which leads into the operating evolution during which the system fulfils its objectives and is modified from time to time to match the minor changes in requirements.
4. Eventually the system becomes less effective, for a number of reasons including mechanical faults, arrival of new technologies and major changes in the requirements and in the environment. This stage leads to decay, which finally leads to replacement of the system – starting at step 1 again.

Liston and Schoene[11] suggest that an effective information retrieval system must have provisions for:

- prompt dissemination of information
- filtering of information
- providing the right amount of information at the right time
- active switching of information
- receiving information in the desired form
- browsing
- getting information in an economical way
- current literature
- providing access to other information systems
- interpersonal communication
- offering personalized help.

These requirements have become even more essential for success in today's web-based information retrieval environment.

Liston and Schoene also talk about how an information retrieval system should be user-oriented, giving primary emphasis to the convenience of the users. Although

this principle has played a key role in the progress of information retrieval, as can be evidenced through the different generations of OPACs and online information retrieval systems, this has become a much more challenging endeavour in the web and digital library environment since the users are remote, and the same information may be required by different users with different characteristics, coming from different locations, cultures and contexts. As a result different approaches and practices may be noticed in different modern information retrieval systems. For example, online database service providers such as Dialog, Ovid, Factiva and CSA have taken various user-centred approaches in their search and retrieval features; examples appear in Chapter 15. Similarly, several specific measures can also be noticed in the information retrieval systems in some digital libraries.

The text retrieval conference (TREC) series of experiments and conferences have been a major platform for advancement of information retrieval research and development activities. Details of TREC experiments, results and commentaries are available on the TREC website and several publications (see for example Voorhees and Harman[12] and Sparck Jones[13]) and these are discussed later in this book.

Information retrieval research and development activities have advanced rapidly over the past few years as a result of the appearance of web and web search engines. Information retrieval, which was once the interest and concern of only a select few – experts and professionals – is now used by everyone for accessing information on the web. Credit must go to web search engines for investing significant amounts of effort and available resources for developing and improving information retrieval systems so that people can get easier and better access to information on the web. Another major contribution of web search engines is the simplification of the search interface and the interaction process. Information retrieval search interfaces that were earlier designed for expert and educated users have now become much simpler and intuitive and can be used by anyone without any specific knowledge of information retrieval techniques or the subject domain. Details of the developments of search interfaces, especially web search interfaces, have been discussed by Hearst.[14]

Discussion

Developments in information retrieval can be viewed from two different perspectives:

- the computer-centred view, which deals with building efficient computer systems for storage, organization and access to information, and focuses on areas such as building up efficient access mechanisms, query processing, ranking algorithms and display and delivery of search results
- the user-centred view, which focuses on the study of human information behaviour, understanding of human needs, information context and use, and so on.

This book aims to focus on both these views of information retrieval, since successful information retrieval systems should take both views into account. It also covers the broader scope of information retrieval that ranges from library OPACs to the web and digital libraries. Alongside the latest developments in computerized information retrieval, this book discusses the traditional library tools and techniques, such as classification, cataloguing and vocabulary control, as well as the traditional manual indexing systems. It is believed that today's information professionals should know about these traditional tools and techniques because of at least two reasons. First, they show the process of the evolution of information retrieval, from the shelf to the web; second, many recent developments in information retrieval in web and digital library environments have their roots, explicitly or implicitly, in traditional bibliographic tools and techniques. In some cases the wheel has been re-invented, perhaps because inventors were not aware of the tools and techniques built and used by libraries over a long period for organization and access to information resources. Major topics covered in this book include:

- organization and processing of information – bibliographic formats, cataloguing, metadata, classification and indexing, vocabulary control
- information retrieval techniques and models
- database and information retrieval systems, online search services
- user interfaces
- information users, human information behaviour and information seeking and retrieval models
- information retrieval evaluation
- markup languages, Hypertext Markup Language (HTML) and XML retrieval
- web information retrieval, digital libraries
- natural language processing, intelligent information retrieval systems
- information retrieval research trends.

References

1 Sparck Jones, K. and Willett, P., Overall Introduction. In Sparck Jones, K. and Willett, P. (eds) *Readings in Information Retrieval*, San Francisco, Morgan Kaufmann Pub. Inc., 1997, 1–7.

2 Parsaye, K., Chignell, M., Khosafian, S. and Wong, H., *Intelligent Databases: object-oriented, deductive hypermedia technologies*, New York, John Wiley, 1989.

3 Lancaster, F. W., *Information Retrieval Systems*, New York, John Wiley, 1968.

4 Belkin, N. J., Anomalous States of Knowledge as a Basis for Information Retrieval, *Canadian Journal of Information Science*, **5**, 1980, 133–43.

5 Meadow, C. T., Boyce, B. R., Kraft, D. H. and Barry, C., *Text Information Retrieval Systems*, 3rd edn, London, Academic Press, 2007.

6 Lancaster, F. W., *Information Retrieval Systems: characteristics, testing, and evaluation*, 2nd edn, New York, John Wiley, 1979.

7 Kent, A., *Information Analysis and Retrieval*, 3rd edn, New York, Becker and Heys, 1971.
8 Vickery, B. C., *Techniques of Information Retrieval*, London, Butterworth, 1970.
9 Vickery, B. and Vickery, A., *Information Science Theory and Practice*, London, Bowker-Saur, 1987.
10 Rowley, J., *The Basics of Information Systems*, 2nd edn, London, Library Association Publishing, 1996.
11 Liston, D. M. and Schoene, M. L., A Systems Approach to the Design of Information Systems. In King, D. W. (ed.) *Key Papers in the Design and Evaluation of Information Systems*, New York, Knowledge Industry, 1978, 327–34.
12 Voorhees, E. and Harman, D. (eds), *Text REtrieval Conference*, Cambridge, MA, MIT Press, 2005.
13 Sparck Jones, K., What's the Value of TREC: is there a gap to jump or a chasm to bridge? *SIGIR Forum*, **40** (1), 2006, 10–20.
14 Hearst, M., *Search User Interfaces*, Cambridge, Cambridge University Press, 2009.