CHAPTER 1

Information Representation and Retrieval: An Overview

Information representation and retrieval (IRR), also known as abstracting and indexing, information searching, and information processing and management, dates back to the second half of the 19th century, when schemes for organizing and accessing knowledge (e.g., the Dewey Decimal Classification of 1876) were created (Wynar, 1980). However, research on IRR did not become a key field in information science until the breakout of World War II. Since then, great minds from various disciplines have been attracted to this emerging field, while information technologies with different degrees of sophistication and maturity are applied to facilitate research and development in IRR.

The history and development of IRR are illustrated here with a focus on the main features of each period and the outstanding pioneers of the field. The key concepts involved in this book are then explained, followed by a discussion of the major components in IRR. This chapter concludes with an exploration of the essential problem of the field, namely, how to obtain the right information for the right user at the right time.

1.1 History and Development of Information Representation and Retrieval

The history of IRR is not long. A retrospective look at the field identifies increased demand, rapid growth, the demystification phase, and the networked era as the four major stages IRR has experienced in its development.

1.1.1 Major Stages

1.1.1.1 Increased Demand (1940s–early 1950s)

World War II denoted the official formation of the field of IRR. Because of the war, a massive number of technical reports and documents were produced to record the research and development activities surrounding weaponry production. Never before had people confronted such an enormous task of IRR, without considering other aspects of information processing and management such as selection, dissemination, and preservation. As Bush (1945) wrote:

> There is a growing mountain of research. But there is increased evidence that we are being bogged down today as specialization extends. The investigator is staggered by the findings and conclusions of thousands of other workers—conclusions which he cannot find time to grasp, much less to remember, as they appear. (p. 101)

Indeed, the need for representing and organizing the vast amount of information became quite obvious and pressing. In the fields of chemistry and biology, for example:

> The biomedical press, for example, has been estimated to publish 2 million papers each year (McCandless, Skweir, & Gordon, 1964, p. 147). These papers can be read at the rate of two per hour—assuming that the reader is attentive, can read approximately 70 languages, and has the documents at hand. If journal reading is limited to 1 hour per day and 365 days per year, then it will take more than 27.4 centuries to read the output of 1 year of the world's biomedical press. (Borko & Bernier, 1975, p. 6)

Although the exact amount of technical information produced during the 1940s and 1950s may not be accurately determined, its magnitude can be estimated according to the previously described biomedical field. People could no longer rely exclusively on their own skills, memories, or individual file cabinets for organizing and retrieving information efficiently whenever there was a need. Rather, collective efforts in the area of IRR were called for, resulting in systems specifically designed for that purpose even though they were manual ones such as the coordinated indexes introduced in 1951 (Gull, 1956).

1.1.1.2 Rapid Growth (1950s–1980s)

The decades in this period represent the golden years in the development of IRR. Computers were formally introduced to the field between 1957 and 1959, when Hans Peter Luhn used one to handle, not only keyword matching and sorting tasks, but also the intellectual work related to the content analysis of written texts (Salton, 1987).

The emergence of online systems such as DIALOG in the 1960s and 1970s signified the shift from manual to computerized information retrieval (IR). Hahn (1996) described the pioneer online systems developed during that time:

> [They] had some remarkable advanced features such as online thesauri, ranked output, automatic inclusion of synonyms in the search formulation, Boolean logic, right and left-hand truncation,

> cited-reference searching, and natural language free-text searching. Some systems had automatic data collection programs to monitor use and satisfaction. (p. 34)

Along with the growth and maturity of the online systems, automated and automatic techniques for IRR were being created and experimented with, supported by advancements in computing technology. People from different fields, especially computer science, devoted their efforts to research and development in this area. However, additional problems were waiting to be researched, just as Salton (1987) summarized in one of his writings:

> Even though great progress has been made in the past 30 years in text processing and information retrieval, especially in the areas of text editing and document production, index term assignment, and dynamic collection search and query formulation, few substantive advances are apparent in the true text understanding areas. (p. 379)

1.1.1.3 Demystification Phase (1980s–1990s)

Although the online systems previously described were built for users with various kinds of information needs, those systems were not designed in such a way that the users could search them directly, without any training or assistance from information professionals. In other words, only intermediaries such as librarians and other information professionals were able to perform the search task on behalf of the users. Moreover, it was very costly to use such systems for locating information because a remarkable array of fees (e.g., telecommunication, connection, and database fees) would be charged for every single search. The term *end users* then referred to those who had some information need but would not conduct the actual online search themselves. The meaning of the term end user changed gradually when personal computers began to be employed in IR and when CD-ROM and online public access catalog (OPAC) systems were implemented in the mid-1980s.

Online IR systems in the past were accessed via assorted means, such as printer terminals and Cathode Ray Tube (CRT) terminals. Needless to say, the interaction between the searcher and the system was not inviting or friendly. When personal computers were introduced into IR, end users found that the retrieval process seemed much less intimidating because some form of friendly "conversation" or interaction could be carried out between the user and the system.

The implementation of CD-ROM and OPAC systems made it possible for end-users to search for themselves by demystifying IR systems that had

existed previously in online form only and were geographically located in remote places, users no longer needed to be concerned with the online costs when they searched CD-ROM or OPAC systems. Since then, IR systems have increasingly become systems built for and used by end users.

1.1.1.4 The Networked Era (1990s–Present)

IR, up to this point, was a centralized activity, meaning that databases of IR systems were physically managed in one central location. If people intended to use several IR systems, they had to establish connections with respective IR systems individually. In comparison, distributed searching allows people to access database information using the network infrastructure. IR systems are no longer restricted to one single geographical area. The advent of the internet truly makes networked IR a reality by providing the infrastructure needed for this implementation.

In addition to the feature of distributed searching, the internet has redefined the field of IRR. Never before in the history of information representation have statistical keywords and similar methods been applied so extensively to such a huge amount of hyperstructure and multimedia information. Never before in the history of IR have so many users conducted online searches without the help of intermediaries. As a result, the quality of information representation in this environment seems so mixed that the term *organized chaos* has been coined specifically for describing the status quo. On the other hand, full-text retrieval has become the norm rather than the exception on the internet. Retrieval techniques that were previously tested only in lab experiments are readily available in internet retrieval systems such as AltaVista and Google. Overall, research results obtained from controlled environments are now widely applied to IRR on the internet.

The networked era is symbolized by the internet, which provides a novel platform as well as a showcase for IRR in the digital age.

1.1.2 Pioneers of the Field

The field of IRR has attracted so many talented and devoted people in the past 50 years that it is impracticable to list all of them here. Nevertheless, the following people deserve individual discussion in this book because their contributions to the field are so great. Another criterion for selecting the IR pioneers in this section is that all of them had concluded their academic careers by the time of this writing.

1.1.2.1 Mortimer Taube (1910–1965)

Mortimer Taube earned his doctorate in philosophy at the University of California at Berkeley. He worked as a librarian for circulation, cataloging,

and acquisitions before he took a position at the Library of Congress to become the assistant chief of general reference and bibliography in 1945 (Shera, 1978). In 1952, Taube founded Documentation Inc., where he and his colleagues began to explore new methods for information indexing and retrieval under contract to the U.S. Armed Services Technical Information Agency (Smith, 1993).

The new approach to indexing and searching eventually became the well-known *coordinate indexing*. Taube, with Alberto F. Thompson, presented a report titled "The Coordinate Indexing of Scientific Fields" before the division of chemical literature of the American Chemical Society as part of its Symposium on Mechanical Aids to Chemical Documentation. The report was never officially published in a journal or book, but Gull (1987) included it later in one of his publications as an appendix.

The need for new indexing and retrieval methods at the time was twofold. First, a huge number of technical reports and other scientific literature, generated by research conducted for World War II, indicated the inadequacy of the existing indexing and retrieval systems, which were primarily manual. Second, the two established methods of representing information, the alphabetical and the hierarchical (e.g., subject headings and classification schemes), were unable to accommodate the new disciplines, new technologies, and new terminology that evolved from research and development related to World War II (Smith, 1993). It was in this particular circumstance that in 1952 Taube and Thompson proposed coordinate indexing (Gull, 1987).

Breaking away from traditional methods for indexing and searching, coordinate indexing is based on the implementation of uniterms and application of Boolean logic in IR. Uniterms are individual terms selected by indexers to represent different facets of a document. Uniterms can, in a sense, be regarded as today's keywords because both are derived from original documents, and no effort is made toward vocabulary control (e.g., checking synonyms and homographs). Typically, several uniterms are used to represent a single document, as with keyword indexing.

Boolean logic, a subdivision in philosophy, was put forward by George Boole in 1849 on the basis of his detailed analyses of the processes of human reasoning and fundamental laws that govern operations of the mind (Smith, 1993). To Boole, the processes of reasoning were either the addition of different concepts, or classes of objects, to form more complex concepts, or the separation of complex concepts into individual, simpler concepts (Boole, 1854). The former is summarized as the AND operator, and the latter includes the OR and NOT operators. Almost a century later, Taube brought these principles to the field of IRR in the form of coordinate indexing.

Taube's cumulative efforts in coordinate indexing laid the foundation for Boolean searching in the computerized environment. If disciplines can be

broken down into single ideas represented by uniterms, computers can be used to organize and search for information put in that format. This insight eventually led to the development of various retrieval systems that performed all types of Boolean searching, a topic explored further in other parts of this book. Meanwhile, indexes that were developed using the coordination method became known as coordinate indexes. The process of searching these indexes by combining ideas to find needed information was called *concept coordination* (Smith, 1993).

The phrase *coordinate indexing* is, however, a misnomer (Gull, 1987). More accurately, it should be called *coordinate indexing and searching* or, in today's terminology, *coordinate representation and retrieval*, because it does not appear to be just an indexing method. Rather, it has been used for searching as well. In addition, the emphasis of the method was put on analysis over synthesis, which led some critics (e.g., Gull, 1987; Pao, 1989) to question whether coordinate indexing combined words or concepts/ideas. The problem of false coordination also caused concern because there was no mechanism in the method to prevent false drops. For example, if the desired search topic is *computer desk*, the retrieval results may contain documents related to *computer desk*, as well as *desk computer* and other phrases that happen to have the words *computer* and *desk* in them (e.g., *desktop computer*). As uniterms are assigned without reference to controlled vocabularies, all the disadvantages associated with indexing and retrieval using natural language (see Chapter 4) are present in coordinate indexing. Furthermore, if coordinate indexing is limited to single words only, it seems just a restriction imposed on the old alphabetical and hierarchical methods rather than a novel third method for indexing and retrieval (Gull, 1987).

Nevertheless, Taube's contribution to the field is remarkable in that he created coordinate indexing and introduced Boolean logic to indexing and retrieval. As summarized by Gull (1956), the uniterm system exhibited the following characteristics, in comparison with classification and subject catalogs:

- Lower cost
- Smaller size
- Faster analysis
- More access points per unit cataloged
- Faster searching
- Slower rate of growth
- Slower rate of obsolescence
- Greater specificity

- Universality
- Logical structure
- Neutrality
- Simplicity
- Suitable for cumulative publication

There is certainly little doubt that the real impetus to modern methods of IR was given by Mortimer Taube, founder of Documentation Inc. (Lancaster, 1968). As we move from the electronic age to the digital one, the influence and impact Taube had on indexing and retrieval remain apparent.

1.1.2.2 Hans Peter Luhn (1896–1964)

If Taube is regarded as the pioneer who laid the foundation for the application of computers in IR by incorporating Boolean logic in his uniterm system, Hans Peter Luhn is the individual who actually created computer-based applications for the field.

Born in Germany, Luhn was an engineer by education who became a famed IBM inventor with more than 80 patents. Luhn's initiation into information science in general and IR in particular came in the period 1947–1948, when James Perry and Malcolm Dyson approached him and asked whether an IBM machine could be designed to search chemical structures coded according to the Dyson notation system (Harvey, 1978). Luhn quickly became interested and joined with them to develop and test a pioneer electronic information searching system that, in 1948, Luhn called an electronic searching selector. This machine later came to be known as the Luhn scanner (Schultz, 1968). By 1953, he was spending an increasing amount of time in IR and published his first paper in the area, titled "A New Method of Recording and Searching Information" (Luhn, 1953). He also became the manager of IR research at IBM. Luhn explored and worked out many of the important computer-based IR applications that now seem commonplace in the field.

One such application is the KeyWord In Context (KWIC) system, which encompasses three elements fundamental to IRR. The first element is that keywords, rather than terms obtained from conventional classifications and subject headings, are used to represent and retrieve the plural facets of a document. Keywords, in a sense, can be considered a descendant of Taube's uniterms, although few people have attempted to make this association. This keyword approach has since been widely implemented in applications such as automatic indexing, automatic abstracting, and keyword searching. The second element of the KWIC method derives from the concordances our

ancestors created as early as the 13th century (Wellisch, 1995). While all the sentences in a document make up a concordance, titles and similar artifacts (e.g., topic sentences) form the so-called context for KWIC products. For instance, titles are processed to generate KeyWord In Title (KWIT) indexes, one extension of the KWIC application. The third element of the KWIC approach is the permutation of keywords contained in titles and other equivalents. The permutation typically assumes two display formats: KWIC with a particular keyword aligned within the context, and KeyWord Out of Context (KWOC) with a particular keyword displayed and left aligned outside the context. KWOC, like KWIT, is a variation of the KWIC application. Luhn, who coined the KWIC terminology in 1958, successfully produced a KWIC index for *Chemical Titles*, bringing his idea to a practical result (Fischer, 1966). The KWIC approach is unarguably a significant milestone in IRR.

Automatic indexing and abstracting represent major contributions made by Luhn to the IR field. Using statistical methods, Luhn developed and promoted algorithms for producing indexes and abstracts automatically. For automatic indexing, the procedure is based mainly on the selection of significant words (i.e., keywords) that carry meanings from documents. Words that appear frequently (i.e., high-frequency words such as articles, conjunctions, and prepositions) or seldom in the document (i.e., low-frequency words or words people rarely use in communication) and *noninforming words* (i.e., general nouns such as *report* and *summary*, as well as terms constantly used in a particular collection, e.g., *information retrieval* in a document database for IR) can be eliminated by adopting a stop-word list or statistical word-frequency procedure (Luhn, 1958). This approach for producing KWIC indexes, KWIC variations (e.g., KWOC and KWIT), and other keyword indexes became the first and—to this day—the only fully automatic indexing method. For automatic abstracting, two measures are suggested for identifying significant words and subsequently significant sentences that can be the most representative of a given document. The keyword approach, as described earlier for automatic indexing, furnishes one of the two measures for constructing automatic abstracts. The other measurement relies on the relative position within a sentence of significant words. According to Luhn (1958), proximity of four or five nonsignificant words between significant words appears useful for selecting significant sentences from a document. A combination of keyword frequency and keyword proximity within a sentence seems a viable methodology for generating auto-abstracts.

Luhn's third notable contribution to the field was the development of selective dissemination of information (SDI) systems. SDI is an application for effective dissemination of new scientific information to target users based on their profiles. Luhn (1959) outlined its components and various steps for

operating an SDI system, of which the creation and maintenance of user profiles is the most important and critical task. The profile of user interests includes a list of words, along with their current weights, each indicating the balance between additions and subtractions resulting from the profile maintenance procedure. The profile is then checked against document representations (e.g., abstracts and index terms) at a specified temporal interval (e.g., weekly or monthly). As conceived by Luhn, intelligent systems could be built for business, science, and other types of literature with the implementation of the SDI concept (Stevens, 1968).

Obviously, Luhn put quite a few ideas into practice for IRR with the aid of computers even though not all of them originated from him (Stevens, 1968; Wellisch, 1995). As all the aforementioned applications are computer-based, the efficiency of those IR-related operations has been drastically improved. However, these applications cannot attain the quality generally associated with indexes, abstracts, and retrieval tasks that are done manually.

Luhn's contributions to the field, especially to computerized IRR, earned him an important place in the history of information science, of which IR is a key component. Based on an analysis by Carlos Cuadra (1964), Luhn's name led all the rest on three out of four of the listings of major contributions as perceived by four experts in information science. Luhn also ranked fourth among the top 25 authors in terms of publication density, a score calculated by Cuadra using bibliographic data. In addition, Luhn was positioned in the top 10 in the four bibliographies of the most frequently cited authors in the field. All the findings confirm that Luhn stands out as a prominent researcher in information science, particularly in IR, regardless of the evaluation method (e.g., expert advice, textbook analysis, and citation analysis) used (Cuadra, 1964). It is Luhn who brought computers into our field, pioneered many IRR applications, and catalyzed empirical research in IRR.

1.1.2.3 Calvin N. Mooers (1919–1994)

Compared with Taube and Luhn, Calvin N. Mooers' contributions to IRR came much later. Mooers' areas of study were mathematics and physics, but he did devote a considerable amount of his time to information and computer science after attending the Massachusetts Institute of Technology in 1946 to capitalize on his computing experience (Corbitt, 1992).

In 1950 Mooers in fact coined the term *information retrieval*, which has since been seamlessly integrated into the vocabulary of information science. According to Mooers, IR means finding information whose location or very existence is a priori unknown (Garfield, 1997). Mooers (1960) was also credited with proposing Mooers' law for IR systems:

> An information retrieval system will tend not to be used whenever it is more painful and troublesome for a customer to have information than for him not to have it. (p. ii)

Mooers' law has been paraphrased, and one version reads, "An information system will only be used when it is more trouble not to use it than it is to use it" (Koenig, 1987). Garfield (1997) further suggested a corollary to Mooers' law: "The more relevant information a retrieval system provides, the more it will be used." In sum, Mooers' law indicates quintessentially that systems that reflect users' needs and practices are more apt to be consulted readily (Henderson, 1996).

In addition to coining the term *information retrieval* and authoring Mooers' law, Mooers developed the zatocoding system for storing a large number of document descriptors on a single, specially notched card by superimposing random, eight-digit descriptor codes. The use of the zatocoding system would result in only a small but tolerable number of false drops in a bibliographic search (Garfield, 1997). Mooers was also responsible for creating two applications oriented toward computer science: the Text Reckoning and Compiling (TRAC) and VXM computer languages. TRAC was designed specifically for handling unstructured text in an interactive mode as opposed to the batch mode. VXM was used for multicomputer network systems (Corbitt, 1992; Henderson, 1996).

In recognition of his outstanding contributions to the field of information science, Mooers was honored with the American Society for Information Science Award of Merit in 1978. The award citation states that he "has affected all who are in the field of information and his early ideas are now incorporated into today's reality" (Henderson, 1996). Indeed, Mooers was one of the great pioneers in the growing field of IRR.

1.1.2.4 Gerard Salton (1927–1995)

Everyone in the IR community agrees that Gerard Salton was one of the preeminent figures in the field. He was the man "most responsible for the establishment, survival, and recognition of IR …" by spending 30 years of the latter part of his life "carefully nurturing it and sustaining it when the professional climate was inhospitable, and defending it until it could support itself" (Crouch, et al., 1996). If the entire field of IR were regarded as a domed architecture, Salton would be the dome, and his colleagues and protégés would serve as pillars or other supporting parts of the structure.

Salton's main research tool was the System for the Manipulation And Retrieval of Texts (SMART), also humorously known as "Salton's magical automatic retriever of text." His ideas fundamentally changed full-text processing methods on computers and provided the field of IR with solid

underpinnings (ACM SIGIR, 1995). "His research contributions span the gamut of information retrieval: the vector space model, term weighting, relevance feedback, clustering, extended Boolean retrieval, term discrimination value, dictionary construction, term dependency, text understanding and structuring, passage retrieval—and, of course, automatic text processing using SMART" (Crouch, et al., 1996). Each of these contributions can easily be a topic for extensive discussion and some are discussed in other parts of this book. Nevertheless, it is not an exaggeration to say that Salton brought computer science and contemporary techniques to IR.

IR systems that operated commercially in the 1960s were basically using Boolean logic and other pre-SMART retrieval technology. Today, dozens of well-known commercial systems use the ideas and technology developed in SMART. For example, Individual (a news clipping service) licensed the SMART technology directly. Others, such as the wide area information servers (WAIS) and DOWQUEST, a tool for the Dow Jones newswire, are directly derived technology. Many new systems have leveraged off the years of research, including WIN, a legal retrieval system run by the West Publishing Company, and INQUERY, another eminent research tool (ACM SIGIR, 1995). IR techniques, previously only tested in SMART, are now commonly implemented, even in the newest species of IR systems: internet retrieval systems.

All these systems show the influence and application of Salton's IR concepts and research in the electronic and digital environments. In addition, Salton was a prolific writer. He published five texts on IR and more than 150 research articles in the field throughout his career (ACM SIGIR, 1995). He also dedicated his outstanding services to the field. In return, Salton received numerous awards.

1.1.2.5 Karen Spärck Jones (1935–2007)

Karen Spärck Jones completed her undergraduate degree first in history and later in philosophy at the University of Cambridge, U.K. It was her education in philosophy that led her to the field of IR, particularly in what is now labeled natural language processing (NLP). In her PhD dissertation in 1964, titled "Synonyms and Semantic Classification," she employed the theory of clumps for term clustering (Wilks, 2007).

Spärck Jones, working in IR for more than four decades, made significant contributions to several areas and influenced many around her, both colleagues and students. Chronologically speaking, the experimental approach to IR was the first of such areas. During the final years of Spärck Jones' dissertation research, the epic Cranfield tests entered their second phase, and Spärck Jones began showing a strong interest not only in the experimental approach Cranfield took but also in the details of experimental methods and

materials (Robertson & Tait, 2008). She consequently used the Cranfield collection in experiments to test term clustering for semantic classification, which she had begun exploring in her doctoral research.

Her experimental research in IR did not stop, however, when the Cranfield tests ended in 1967. Rather, she led a joint effort of U.K.-based researchers in the 1970s to develop a new "ideal" test collection so that experiments could be carried out beyond a single one like the Cranfield tests. Although this effort was not brought to fruition due to the insufficient financial support available in the U.K., she continued toiling tirelessly in this area ever. Spärck Jones edited *Information Retrieval Experiment* (Spärck Jones, 1981), the only monograph devoted primarily to experimental methods in IR. She also wrote two of the papers in that book. On the other hand, when the Text REtrieval Conference (TREC) series commenced in 1992, Spärck Jones had attained the ideal test collection she and her colleagues had been unable to create almost two decades earlier. From the very beginning, she enthusiastically participated in the TREC series in various capacities (e.g., informal advisor to the organizers, program committee member, and research team participant). More important, Spärck Jones authored a series of papers summarizing and comparing different teams' performances in the TREC experiments. She also reflected on the lessons the IR community could learn from the multitude of disparate results by TREC participants (Spärck Jones, 1995, 2000). No one seemed more appropriate than Spärck Jones to undertake this tremendously significant task, given her lifelong interest in and commitment to the experimental approach to IR.

The paper Spärck Jones published in 1972 on inverse document frequency (idf) represented another area of her research: statistical methods in IR (Spärck Jones, 1972). She explained in that paper, still one of the most highly cited papers in the field, that a document is relevant not only because key terms are frequent in it but because those terms are infrequent in other nonrelevant documents. In conjunction with the term frequency (tf) algorithm developed by Salton's group at Cornell University, the tf.idf combination became the most common default weighting scheme for many years. The idf criterion further led to relevance weighting in the probabilistic model Robertson developed in collaboration with Spärck Jones (Robertson & Spärck Jones, 1976). In a sense, Spärck Jones' work on idf inspired a series of well-founded experimental investigations of term weighting. The statistical methods, exemplified by the idf and other measures of similar nature, have proven to be inexpensive and competitive in IR, and the most recent testimony can be seen in the TREC series (Spärck Jones, 1995).

Extending her research on statistical methods in IR into a larger and different but closely related area, Spärck Jones started work in the 1980s on natural

language processing, with a specific focus on automatic summarization, question answering, and natural language querying. She strongly believed that "words stand only for themselves" (Wilks, 2007) and found it strange to represent text using languages (e.g., subject headings) other than the natural language (Spärck Jones, 1994). Her research on natural language processing (e.g., Spärck Jones, 2005, 2007) eventually built her a reputation at least as significant in that area as in others, although NLP is different from them in comparison. Because of Spärck Jones' important contribution to natural language processing, a book titled *Charting a New Course: Natural Language Processing and Information Retrieval* (Tait, 2005) was published in her honor. This book is not the only publication that manifests her contribution and influence in NLP. In fact, in addition to many other publications she authored, she jointly edited a reader in NLP as early as 1986 (Grosz, Spärck Jones, & Webber).

Spärck Jones' research in later years involved spoken document retrieval. With colleagues, she authored two award-winning papers about their innovative work on this topic (Maybury, 2005). Robust unrestricted keyword-spotting algorithms and adapted, existing text-based information retrieval techniques were among the research Spärck Jones conducted on voice data such as speeches and broadcast news.

The IR field has benefited enormously from Spärck Jones' work in the idf weighting method, natural language processing, spoken document retrieval, and the experimental approach to IR (Willett & Robertson, 2007). Her influence on the IR community as a researcher, mentor, and advisor will continue to be felt for many years.

1.2 Elaboration on Key Concepts

The title of this book contains four distinct concepts: information, information representation, IR, and digital age. Each of the concepts, without exception, has synonyms and can be interpreted and understood differently in different contexts. The following discussion is intended to explain and clarify the meaning and connotation of the concepts in this book.

1.2.1 Information

Information, as a concept, has been considered and contrasted extensively with such terms as *data*, *knowledge*, and *wisdom* (e.g., Meadow, 1992). Hence, there seems little need for this book to repeat or continue those discussions or debates. On the other hand, the words *information*, *text*, and *document* are often used interchangeably in the field (e.g., *text retrieval* and *document representation*). According to Larsen (1999):

> In our field, documents are characterized by having a "price," they can be counted in "numbers," and are in that capacity the basic components of library statistics. Most of them "take up space," they can be "damaged in use," and they may "deteriorate" over time. (p. 1020)

In addition to Larsen's definition, documents may contain multimedia. If *text* refers to textual information only, *document* could include multimedia information (i.e., any combination of audio, video, image, and textual information). It appears that *information* encompasses both text and document, having the broadest connotation among the trio.

In recent years, research has been done on passage retrieval as opposed to document retrieval (Spärck Jones, 2000). *Passage retrieval* (also called information retrieval) denotes finding the very information or document passage (e.g., a paragraph or an arbitrary length of document segments) the end user needs. In contrast, *document retrieval* implies getting a full document for the end user, even if only one short passage is needed. If the word *information* is treated as a synonym of *passage*, as in *passage retrieval*, one exception must be made to the previous discussion regarding the implication of *information*.

1.2.2 Information Representation

No matter which format a piece of information may take, it needs to be represented before it can be retrieved. Information representation includes the extraction of some elements (e.g., keywords or phrases) from a document or the assignment of terms (e.g., descriptors or subject headings) to a document so that its essence can be characterized and presented. Typically, information representation can be done via any combination of the following means: abstracting, indexing, categorization, summarization, and extraction. *Information processing* and *information management*, though having different meanings, are often regarded as synonyms of *information representation*. While *information processing* refers to how information is handled for retrieval purposes, *information management* deals with the full range of activities associated with information, from information selection to information preservation.

In this book, the term *information representation* will be used to cover the various aspects and methods of creating surrogates or representations (e.g., indexes and abstracts) for IR purposes.

1.2.3 Information Retrieval

Information retrieval has been treated, by and large, as a subject field covering both the representation and retrieval sides of information (Spärck Jones & Willett, 1997). The retrieval dimension is further referred to as *information access*, *information seeking*, and *information searching*. These terms can be considered as synonyms for *retrieval*. However, each of them does have its different orientation with regard to implications. The term *information access* emphasizes the aspect of getting or obtaining information. In contrast, *information seeking* focuses on the user who is actively involved in the process. As for *information searching*, the center of attention appears to be on how to look for information.

In addition to the aforementioned terms, *data mining* and *resource discovery* have often been found in the information professional's vocabulary in recent years when IR is discussed. Both terms are normally used in the business and networked environment. At this time, it remains to be seen whether *data mining* and *resource discovery* will be incorporated into the permanent vocabulary of people from the field of IR.

Another layer of meaning in IR is *information storage*, which mainly deals with the recording and storage of information. However, such usage is gradually becoming a past practice because *information storage* is no longer a major concern, thanks to the advances made in information storage and access technology. Therefore, IR in this book encompasses information seeking, information searching, and information access but excludes information storage.

1.2.4 Digital Age

The word *digital*, as opposed to *analo*g, is a relatively new concept. Both *digital* and *analog* are terms related to electronic technology, according to the following account provided by Tech Target (2008), a company also defining terminology in information technology:

> Digital describes electronic technology that generates, stores, and processes data in terms of two states: positive and non-positive. Positive is expressed or represented by the number 1 and non-positive by the number 0. Thus, data transmitted or stored with digital technology is expressed as a string of 0's and 1's. … Prior to digital technology, electronic transmission was limited to analog technology, which conveys data as electronic signals of varying frequency or amplitude that are added to carrier waves of a given frequency. Broadcast and phone transmission has conventionally used analog technology.

With the advent of the computer, the internet, and other information technologies, we have apparently been entering into a digital age. Various activities, including research and development, related to IRR increasingly take place in the digital environment. This book is thus titled accordingly, signifying the state of the art for the field.

1.3 Major Components

IRR in its totality can be divided into several major constituents: database, search mechanism, language, and interface. People (including the user, the information professional, and the system developer), information, and systems are the three intertwining entities that function jointly in the process of IRR although they are not discussed specifically in this section.

1.3.1 The Database

Databases in IRR comprise information represented and organized in a certain manner. In the traditional sense, a database (e.g., an online database) is typically made of records that can be further decomposed into fields, the smallest and most natural units for sorting, searching, and retrieving information. In a database of journal publications, for instance, author and title are two fields. Traditional databases consist of two parts: *sequential files* and *inverted files*. The sequential file is the database source, containing information organized in the field-record-database structure. It is called a *sequential* file because the records in it are ordered according to the sequence they take when being entered into the database. The inverted file, also known as an index, provides access to the sequential file according to given search queries. It is called an *inverted* file because the order in which information is presented (access point first and locator second) is the reverse of that in sequential files (locator first and access point second).

In a nontraditional sense, a database (e.g., as in internet retrieval systems) may still have sequential and inverted files. But the composition of the sequential file for an internet retrieval system, for example, is different from its counterpart for traditional IR systems (e.g., online systems) in that the composition of the sequential file does not take the field-record-database structure. Rather, the sequential file is made of fieldless information presented in proselike format. In addition, information contained within is not a surrogate (e.g., abstract) but a part or the full content of an original internet document (e.g., a webpage). In traditional IR systems, however, information in sequential files is usually some type of representation in the form of bibliographical descriptions, abstracts, summaries, extracts, and the like.

The content and coverage of the database determine what can be retrieved later from the IR system.

1.3.2 The Search Mechanism

Information represented and organized systematically in a database can be searched and retrieved only when a corresponding search mechanism is provided. A search mechanism can acquire any degree of sophistication in search capabilities, which are defined ultimately by the search algorithms and procedures the IR system incorporates. All search procedures can be categorized as either basic or advanced. Basic search procedures are commonly found in the majority of operational IR systems, while advanced search procedures have been tested and experimented with mainly in laboratories or prototype systems. However, in recent years, advanced search algorithms have increasingly been integrated into internet retrieval systems.

Procedures such as keyword searching, Boolean searching, truncation, and proximity searching belong to the basic search algorithm cluster. As noted earlier, these procedures often form the search mechanism embedded in many IR systems. End users with little training or search experience should be able to perform simple searching tasks in retrieval environments of this kind. Advanced search procedures, such as weighted searching, are employed mostly in newer retrieval systems and are generally designed for people with professional training and search experience.

The capacity of a search mechanism determines what retrieval techniques will be available to users and how information stored in databases can be retrieved.

1.3.3 The Language

Information relies on language, spoken or written, when being processed, transferred, or communicated. In this context, language is one of the crucial components in IRR. Language in IRR can be identified as either natural language or controlled vocabulary. What people naturally use for representing information or forming a query is called *natural language*. If an artificial language, whose vocabulary, syntax, semantics, and pragmatics are limited, is applied in IRR, that language is termed *controlled vocabulary* (Wellisch, 1995). There are three common types of controlled vocabulary: classifications, subject heading lists, and thesauri, each with its own special usage in IRR.

Natural language, generally speaking, allows the highest degree of specificity and flexibility in representing and retrieving information. People do not need any training or practice in using natural language because it is what they use for oral and written communication every day. In contrast, it is

costly to create and maintain a controlled vocabulary, and people have to learn how to use it by practice and training. Nevertheless, controlled vocabulary is able to reduce the intrinsic difficulties (e.g., complexity, subtleness, and ambiguity) in using natural language for representing and retrieving information (Lansdale & Ormerod, 1994). The debate in this field about natural language versus controlled vocabulary has been going on since the end of the 19th century (Rowley, 1994). However, language remains an essential part of IRR regardless of debate outcomes.

Language, to a certain degree, determines the flexibility and artificiality in information representation and retrieval. Chapter 4 of this book is devoted to the discussion of language in information representation and retrieval.

1.3.4 The Interface

Interface, according to Shaw (1991), is what the user sees, hears, and touches while interacting with a computer system. In IRR, interface refers to the interaction occurring between the user and related activities. Also, with this component of IRR, the user dimension appears obvious yet is intermingled with the other three components: database, search mechanism, and language.

Interface is regularly considered when judging whether an IRR system is user-friendly. As defined by Mooers' law, user-friendly systems will attract more people than user-hostile ones in terms of usage. The quality of an interface is decided by interaction mode (e.g., menu selection), display features (e.g., screen layout and font type), and other related factors. Adaptive and affective technologies are beginning to be applied in interface design and implementation as more attention is paid to the human dimension in IRR. Interface determines the ultimate success of a system for IRR, especially if the system operates in the digital environment.

In sum, database, search mechanism, language, and interface constitute the major components of IRR that interact with the human dimension at one stage or another during the IRR process.

1.4 The Essential Problem in Information Representation and Retrieval

The essential problem in IRR remains how to obtain the right information for the right user at the right time despite the existence of other variables (e.g., user characteristics or database coverage) in the IRR environment. Before exploring the essential problem any further, we should first consider the process of IRR.

1.4.1 The Process of Information Representation and Retrieval

In the IRR process, the user initiates the search and receives any results retrieved, while the information professional is responsible for designing, implementing, and maintaining the IRR systems. Figure 1.1 illustrates the IRR process.

Any information retrieved from the database must first be represented by the information professional according to the language chosen for IRR. Discrepancies are likely to occur during the course of information representation and can be serious problems if a controlled vocabulary is used for the following reasons: First, when information recorded in forms such as journal articles or technical reports is represented as abstracts, indexing terms, and the like, a genuine rendering of the original information does not seem achievable. One could argue that we represent, for example, a big circle with a smaller one. Even so, the size dimension has been distorted. Second, any controlled vocabulary is merely a subset of the natural language with which the original documents were created. It is therefore often hard to find, for example, an exact match between a term in a document and a descriptor from a thesaurus. For representation purposes, the indexer then has to

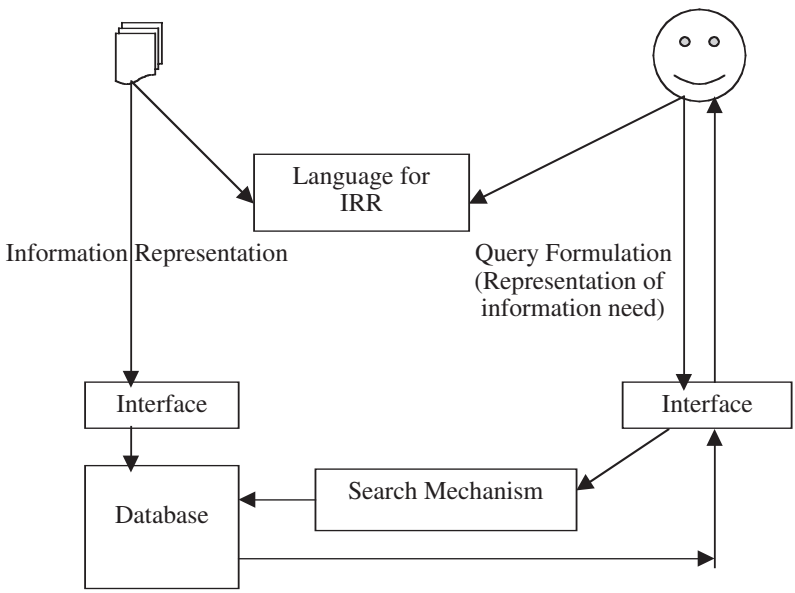

Figure 1.1 Process of Information Representation and Retrieval

choose from related terms, narrower terms, or broader terms listed in the thesaurus. Third, inconsistency in information representation (including concept analysis) appears inevitable, especially if more than one person or system handles the task. Cleverdon (1984) reported that even two experienced indexers, using the same controlled vocabulary, could assign only 30 percent of terms in common to the same document. Similarly, Humphrey (1992) found that inter-indexer consistency in MEDLINE by selecting terms from Medical Subject Headings (MeSH) was less than 49 percent.

On the other hand, users are required to transform their information needs, using the chosen IRR language, into queries that can then be executed in the database with the searching mechanism provided. Researchers have long been aware of the complexity involved in this task. For instance, Blair and Maron (1985) pointed out, "It is impossibly difficult for users to predict the exact words, word combinations, and phrases that are used by all (or most) relevant documents" (p. 295).

In addition, any use of controlled vocabularies and search features (e.g., Boolean operators) will add to the difficulty. Natural language searching, that is, searching with complete phrases or sentences as used in everyday communication without forming any structured queries (e.g., why is the sky blue?), is becoming available on the internet, but there is still a long way to go before researchers in natural language processing—a subdivision of artificial intelligence (AI)—make substantial breakthroughs in their endeavors.

In other words, whether a search will be successful or not depends solely on whether a match is found between the represented information in the system and a query submitted by the user. To be more specific, a search is successful if a match is made between a query and the information represented in the database chosen for the task. Otherwise, a search cannot turn up any useful results. Matching is, therefore, the fundamental mechanism in IRR. As shown in Figure 1.1, there are several points in the IRR process that can cause discrepancies in matching. The ultimate goal for quality IRR is, through the use of various methods and techniques, to minimize or even eliminate the discrepancies that can occur during the process. These methods and techniques are discussed at length in later chapters.

1.4.2 The Limits of Information Representation and Retrieval

While a great amount of research has been done on IRR, there are certain limits that appear insurmountable. When Swanson (1988) pondered the prospect for automatic indexing and retrieval, he borrowed the term *postulates of impotence* (PIs) from E. Taylor Whittaker and formulated nine PIs stating things that cannot be done in IRR. Although made in 1998 with reference

to the automatic domain, several of the statements still seem quite relevant and are quoted here:

> PI 1: An information need cannot be fully expressed as a search request. … The question cannot be precisely formed until the answer is found.
>
> PI 3: … Relevance is not fixed, it is judged within a shifting framework.
>
> PI 4: It is never possible to verify whether all documents relevant to any request have been found. …
>
> PI 9: In sum, the first eight postulates imply that consistently effective fully automatic indexing and retrieval is not possible. The conceptual problems of IR—the problems of meaning—are no less profound than thinking or any other forms of intelligent behavior. (p. 96)

The conceptual problems of IR, as Swanson put it, are critical to the understanding and development of the field. With a close examination of the IRR process, it is apparent that what IRR implies, as suggested earlier, is essentially term matching rather than concept searching in the digital environment. When the search term is, for example, *public transportation*, documents indexed under *buses* or *subway* are not likely to be retrieved unless cross-references are made in the controlled vocabulary. Will IRR one day move beyond term matching and closer to concept searching? The answer is being sought and explored vigorously via trial and error (Swanson, 1977).

References

ACM SIGIR. (1995). Gerard Salton: In memoriam. *IRList Digest*, 12(34). Retrieved January 19, 2009, from www.cs.virginia.edu/~clv2m/salton.txt

Blair, David C., and Maron, M. E. (1985). An evaluation of retrieval effectiveness for a full-text document-retrieval system. *Communications of the ACM*, 28(3), 289–299.

Boole, George. (1854). *An investigation into laws of thought, on which are founded the mathematical theories of logic and probabilities*. London: Walton and Maberley.

Borko, Harold, and Bernier, Charles L. (1975). *Abstracting concepts and methods*. New York: Academic Press.

Bush, Vannevar. (1945). As we may think. *Atlantic Monthly*, 176(1), 101–108.

Cleverdon, C. W. (1984). Optimizing convenient online access to bibliographic databases. *Information Services and Use*, 4, 37–47.

Corbitt, Kevin D. (1992). *Calvin N. Mooers Papers, 1930–1978* (CBI 81). Minneapolis: Center for the History of Computing, Charles Babbage Institute, University of Minnesota. Retrieved October 3, 2009, from www.libsci.sc.edu/bob/isp/mooers.htm.

Crouch, Carolyn, et al. (1996). In memoriam: Gerard Salton, March 8, 1927–August 28, 1995. *Journal of the American Society for Information Science*, 47(2), 108–115.

Cuadra, Carlos A. (1964). Identifying key contributions to information science. *American Documentation*, 15(4), 289–295.

Fischer, M. (1966). The KWIC index concept: A retrospective view. *American Documentation*, 17(2), 57–70.

Garfield, Eugene. (1997). A tribute to Calvin N. Mooers, a pioneer of information retrieval. *The Scientist*, 11(6), 9.

Grosz, Barbara J., Spärck Jones, Karen, and Webber, Bonnie Lynn (Eds.). (1986). *Readings in natural language processing*. Los Altos, CA: Morgan Kaufmann.

Gull, Cloyd Dake. (1956). Seven years of work on the organization of materials in the special library. *American Documentation*, 7(1–5), 320–329.

Gull, Cloyd Dake. (1987). Historical note: Information science and technology: From coordinate indexing to the global brain. *Journal of the American Society for Information Science*, 38(5), 338–366.

Hahn, Trudi Bellardo. (1996). Pioneers of the online age. *Information Processing and Management*, 32(1), 33–48.

Harvey, John F. (1978). Luhn, Hans Peter (1896–1964). In Bohdan S. Wynar (Ed.), *Dictionary of American library biography* (pp. 324–326). Littleton, CO: Libraries Unlimited.

Henderson, Madeline M. (1996). In Memoriam: Calvin N. Mooers, October 24, 1919–December 1, 1994. *Journal of the American Society for Information Science*, 47(9), 659–661.

Humphrey, Susanne M. (1992). Indexing biomedical documents: From thesaurai to knowledge-based retrieval systems. *Artificial Intelligence in Medicine*, 4, 343–371.

Koenig, Michael E. D. (1987). The convergence of Moore's/Mooers' Laws. *Information Processing & Management*, 23(6), 583–592.

Lancaster, F. Wilfrid. (1968). *Information retrieval systems: Characteristics, testing and evaluation*. New York: Wiley.

Lansdale, Mark W., and Ormerod, Thomas C. (1994). *Understanding interfaces: A handbook of human-computer dialogue*. London: Academic Press.

Larsen, Poul Steen. (1999). Books and bytes: Preserving documents for prosperity. *Journal of the American Society for Information Science*, 50(11), 1020–1027.

Luhn, H. P. (1953). A new method of recording and searching information. *American Documentation*, 4(1), 14–16.

Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2(2), 159–165.

Luhn, H. P. (1959). Selective dissemination of new scientific information with the aid of electronic processing equipment. *American Documentation*, 12(2), 131–138.

Maybury, Mark T. (2005). Karen Spärck Jones. In John I. Tait (Ed.), (2005). *Charting a new course: Natural language processing and information retrieval* (pp. xi–xxiii). New York: Springer.

McCandless, R. F. J., Skweir, E. A., and Gordon, M. (1964). Secondary journals in chemical and biological fields. *Journal of Chemical Documentation*, 4(2), 147–153.

Meadow, Charles T. (1992). *Text information retrieval systems*. San Diego: Academic Press.

Mooers, Calvin N. (1960). Mooers' Law: Or, why some retrieval systems are used and others are not. *American Documentation*, 11(3), ii.

Pao, Miranda Lee. (1989). *Concepts of information retrieval*. Englewood, CO: Libraries Unlimited.

Robertson, Stephen, and Spärck Jones, Karen. (1976). Relevance weighting of search terms. *Journal of the American Society for Information Science & Technology*, 27(3), 129–146.

Robertson, Stephen, and Tait, John. (2008). In memoriam: Karen Spärck Jones. *Journal of the American Society for Information Science & Technology*, 59(5), 852–854.

Rowley, Jennifer. (1994). The controlled versus natural indexing languages debate revisited: A perspective on information retrieval practice and research. *Journal of Information Science*, 20(2), 108–119.

Salton, Gerard. (1987). Historical notes: The past thirty years in information retrieval. *Journal of the American Society for Information Science*, 38(5), 375–380.

Schultz, Claire K. (Ed.). (1968). *H. P. Luhn: Pioneer of information science: Selected works*. New York: Spartan Books.

Shaw, Debora. (1991). The human-computer interface for information retrieval. *Annual Review of Information Science and Technology*, 26, 155–195.

Shera, Jesse H. (1978). Taube, Mortimer (1910–1965). In Bohdan S. Wynar (Ed.), *Dictionary of American library biography* (pp. 512–513). Littleton, CO: Libraries Unlimited.

Smith, Elizabeth S. (1993). On the shoulders of giants: From Boole to Shannon to Taube: The origins and development of computerized information from the mid-19th century to the present. *Information Technology and Libraries*, 12(2), 217–226.

Spärck Jones, Karen. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1), 11–21.

Spärck Jones, Karen (Ed.). (1981). *Information retrieval experiment*. London: Butterworths.

Spärck Jones, Karen. (1994). *Finding the information wood in the natural language tree* [Videotape]. Talk presented at the Grace Hopper Celebration of Women in Computing meeting. 41 min.

Spärck Jones, Karen. (1995). Reflection on TREC. *Information Processing & Management*, 31(3), 291–314.

Spärck Jones, Karen. (2000). Further reflections on TREC. *Information Processing & Management*, 36(1), 37–85.

Spärck Jones, Karen. (2005). Some points in a time. *Computational Linguistics*, 31(1), 1–14.

Spärck Jones, Karen. (2007). Automatic summarising: The state of the art. *Information Processing & Management*, 43(6), 1449–1481.

Spärck Jones, Karen, and Willett, Peter (Eds.). (1997). *Readings in information retrieval*. San Francisco: Morgan Kaufmann.

Stevens, Mary Elizabeth. (1968). H. P. Luhn, Information scientist. In Claire K. Schultz (Ed.), *H. P. Luhn: Pioneer of information science: Selected works* (pp. 24–30). New York: Spartan Books.

Swanson, Don R. (1977). Information retrieval as a trial-and-error process. *Library Quarterly*, 47(2), 128–148.

Swanson, Don R. (1988). Historical note: Information retrieval and the future of an illusion. *Journal of the American Society for Information Science*, 49(2), 92–98.

Tait, John I. (Ed.). (2005). *Charting a new course: Natural language processing and information retrieval*. New York: Springer.

Tech Target. (2001). Digital. Retrieved December 3, 2008, from whatis.techtarget.com

Wellisch, Hans H. (1995). *Indexing from A to Z*. 2nd ed. New York: H.W. Wilson.

Wilks, Yorick. (2007). In memoriam: Karen Spärck Jones (1935–2007). *IEEE Intelligent Systems*, 22(3), 8–9.

Willett, Peter, and Robertson, Stephen. (2007). In memoriam: Karen Spärck Jones. *Journal of Documentation*, 63(5), 605–608.

Wynar, Bohdan S. (1980). *Introduction to cataloguing and classification*. 6th ed. Littleton, CO: Libraries Unlimited.