Finding Out About

A Cognitive Perspective on Search Engine Technology and the WWW

RICHARD K. BELEW
University of California, San Diego

CAMBRIDGE
UNIVERSITY PRESS

learning." The chapter ends with "But as AI has moved from a concern with manually constructed knowledge representations to machine learning, and as IR has begun to consider how indexing structures can change with use, these two methodologies have increasingly overlapped."

The last chapter, "Conclusions and Future Directions," reaches out into the future and makes good bedside reading.

C. J. "Keith" van Rijsbergen
University of Glasgow

Preface

One of the things you learn from students is jokes:

> This guy is shopping in a grocery store in Cambridge, Mass. He finishes and lines up with a full basket under a big sign marking the aisle, "Express: 10 items or less." When he gets to the front of the line, the exasperated clerk says, "Look fella, I don't know if you're from Harvard and can't count, or MIT and can't read, but either way you're in the wrong line."

You probably have to have gone to school in Cambridge to really appreciate this joke. I never did, but I find it funny because it laughs at an important division between thoughtful people.

Two Cultures

According to C. P. Snow, the world seems eternally divided into "two cultures, the 'literary intellectuals' and the 'scientists'" [Snow, 1961, p. 4]. Snow himself provides very few clues as to just how we might identify someone at one pole or the other. He suggests that the literate have the unfortunate tendency of falling into the "moral trap":

> . . . which comes through . . . insight into man's loneliness. It tempts one to sit back, complacent in one's unique tragedy, and let the others go without a meal.

Testable hypothesis

He sees scientists and engineers, on the other hand, as optimistic, impatient do-ers! This leads him to hypothesize that "literature changes more slowly than science" (p. 9).† He also thought that, due to the forces of a "fanatical belief in educational specialization" and a "tendency to let our social forms crystallize" (p. 18), the gap between the two cultures was ". . . much less bridgeable among the young than it was even 30 years ago." He said this in 1959! Certainly these same forces have not helped matters in the intervening 40 years.

But Snow's most important recommendation remains true:

> There's only one way out of all this: it is, of course, by rethinking our education . . . . [Speaking especially of British education:] Somehow we have let ourselves the task of producing even a tiny elite educated in one academic skill. For 150 years in Cambridge it was mathematics: then it was mathematics or classics: then natural science was allowed in. But still the choice had to be a single one. (pp. 19, 21)

The premise of this text is that Finding Out About (FOA), the process of actively seeking out information relevant to a topic of interest, absolutely demands a wide-ranging attack by both literary and scientific disciplines. The kind of fractionation that Snow describes has boxed investigators from various disciplines into corners from which they each attempt to address a broad range of fundamentally interdisciplinary questions of *cognition*. FOA is only one such question, but the tension between computational and linguistic sensibilities has been manifest in this domain for an especially long time.

For example, as part of an early meeting of cyberneticists exploring the way that communication and computation might interact, Benoit Mandelbrot, an eminent mathematician and physicist (now most famous for his fractal landscapes), presented hypothetical models of language use that would explain a phenomeon known as **Zipf's law** (a topic discussed in this text, cf. Sections 3.2 and 5.1), claiming these models were analogous to *physical* systems with which he was familiar.

In reaction, A. S. C. Ross, a famous linguist of the 1950s, offered the following commentary:

> [Mandelbrot] states that 'language is a message intentionally produced in order to be decoded word-by-word'. Many schools of linguistic scholarship would reject such a view . . . .
>
> It is, indeed, important that there should be liaison between specialists in communications theory and philologists [linguists, esp. concerned with literature]. The gap between the two subjects is very wide, especially in matters of technique and one wonders what philologists are going to make of remarks such as 'Our model of language is fully analogous to the perfect gas of thermodynamics.'. . .
>
> [But] all statements of this kind really imply that the occurrence of a word at a given point in a text is a matter of chance and this is what philologists and students of literature will deny. If an English writer has to express the idea of TEAPOT – and whether he has to or not is not in the least a matter of chance – the probability of his using the word TEAPOT is unity and the probability of his using the word KETTLE is zero. . . . [Ross, 1953; special typeface not in original]

Mandelbrot's probabilistic models and statistics did not have much to say to at least this linguist.*

An optimist, however, could see a basic *complementarity* between statistical methods and the linguists' syntactic methods. FOA's statistical methods are good at semantics, knowing gross things about an entire document's meaning – what words mean in terms of how they relate to other documents in the corpus and to users' queries. It blithely throws away **noise words** like AND, OF, and THE, because they are assumed to say little about the document's content. Syntactic analysis captures the fine structure of individual sentences and depends critically on the same noise words to reliably anchor its parsing.†

Corpus-based linguistics

The title of this textbook also makes cognitive aspirations. "Cognitive" stems from the Latin *cognitio*, referring to structure, building. We typically imagine cognitive structures to be within an individual's head.

* Ross's comments are reminiscent of van Rijsbergen, p. 127 concerning the probability ranking principle!

But part of what is now known as the discipline of cognitive science is the realization that these representations can be built by many individuals as well as by one. Considering the World Wide Web (WWW) as a representation of knowledge is a topic considered further in Section 6.9.

I am personally drawn to the FOA problem because of the way it intermixes verbal and numeric sensibilities. To say that "literary intellectuals" are interested in language is almost tautological. But one of the major arguments put forward by this text is that many linguistic phenomena also have interesting statistical and mathematical properties. Computations involving these numbers are not only central to the engineering of effective search engines, but they portend fundamental insights into the new forms of communication emerging on the WWW.

Depending on your particular background, some of the techniques and perspectives discussed in this text will come naturally to you, and others will seem as if they are from a different planet. But if you apply some effort at understanding these foreign objects, you may just find out you have lots of new friends in the rest of the solar system. Literate people can learn new mathematical names to apply to their literature, and mathematicians can appreciate new features of the language going on about them.

Typographic Conventions

Other authors who have attempted to discuss language, of course using language to do so, have recognized the confusion that can result as words are used in these two very different roles. Like many of them, I have chosen to use typography to help make this distinction. For example, many of the examples used throughout the text will be drawn from the area of ARTIFICIAL INTELLIGENCE, a subdiscipline of computer science. Terms like this, which are used as examples of lexical items rather than as part of the discourse between me (the author) and you (the reader), will appear as CAPITALIZED and in MONOSPACE FONT.

Second, **boldface** type will be used to flag especially important terms that help to define the FOA problem. For example, **domain of discourse** is the technical term used to describe ARTIFICIAL INTELLIGENCE, the subject matter of the documents we hope to find. These are collected at the end of each chapter, for purposes of review.

Third, the fundamental relation between something in the world and what we think it means is a pivotal issue of this book. But about-ness is also a natural, ubiquitous part of much of our communication, so much so that we will adopt the typographic convention of underlining words such as about and meaning in order to highlight and better appreciate their use.

Finally, authors are always faced with decisions as to which thing they must say first. Making the right decision keeps the story moving forward, while interjecting a digression can make readers lose their way. The WWW is most people's first experience with the **hypertext** alternative to this linear flow. Readers are given the choice points and the opportunity to construct their own *nonlinear* path through a text simply by clicking on links. Obviously such jumps are more difficult to accomplish in a printed text. In this text marginal notes[†] are used to point to tangential topics that a reader might choose to pursue. On the accompanying CD, clicking on the correlated anchor will lead to a brief discussion of this topic. Extra details or clarifications will be provided by footnotes, which are called out in text by asterisks.[*] Traditional numbered footnotes will be used to give URLs of Web sites discussed in the text.

Marginal notes

Audiences

My interest in the topics discussed here goes back to my own dissertation. At that point I was primarily interested in machine learning techniques, and I learned just enough about free-text information retrieval to use it as a demonstration "domain" for the "connectionist" learning techniques I proposed (cf. Section 6.5.2). Since then, I have become increasingly interested in the issues surrounding FOA and have taught courses in Information Retrieval (IR) for many years, at the University of California in San Diego and the University of Wisconsin in Madison.

This book began as a series of lecture notes for these classes. In the first years, I used Keith van Rijsbergen's seminal text [van Rijsbergen, 1979]. (This book was already out of print when I first found it, but ***van Rijsbergen's text***[†] has now been placed in its entirety on the WWW.)

[*] Footnote.

[†] http://www.dcs.gla.ac.uk/~iain/keith/

This text so influenced my thinking on this subject that it occupies a special relationship with *FOA*: I quote from it especially often, and I use the special referential convention of van Rijsbergen, p. iii. With Keith's permission, I include a complete copy of his hypertext on the FOA CD, and every reference to that text will allow you to click and go directly to the cited page.

Several other texts deserve special mention. The collection of chapters edited by Frakes and Baeza-Yates [Frakes and Baeza-Yates, 1992] provides an excellent introduction to many topics; Fox's chapter 7 in particular figures heavily in Chapter 2 of this text. Baeza-Yates and Ribeiro have recently edited another collection of very useful chapters [Baeza-Yates and Ribeiro, 1999]. As I was finishing work on this book, Manning and Schütze produced an excellent survey of corpus-based linguistic techniques [Manning and Schütze, 1999] that extends significantly beyond the basics provided in Section 6.3.2. Robert Korfhage has written a textbook that is especially useful from the perspective of library science [Korfhage, 1997]. I highly recommend *Readings in Information Retrieval*, edited by Karen Sparck Jones and Peter Willett [Sparck Jones and Willett, 1997], as a companion to this text. That collection pulls together many classic papers from IR's distant past, some of which are now hard to get. A supplement (available at the ***FOA Web site***[2]) links readings from that text as an adjunct to this textbook.

Because I teach primarily in a Computer Science department, the primary audience for this textbook is computer science students, both graduate and undergraduate, like those I have had the good fortune to meet in my classes. At the same time, I have tried to suppress technical details or explain them in ways that should make the most important themes accessible to audiences (e.g., linguists, library scientists) who are more comfortable with words than with equations. Search engine technologies are central to the FOA problem, but this text was designed to be accessible to those who write such computer programs as well as to those who do not.

Executable versions of all basic routines are available on the attached CD-ROM; current versions are maintained at the ***FOA Web site***[2]. Together with the test corpora and experimental data (queries, relevance assessments), students and teachers should be able to explore many

[2] http://www.cs.ucsd.edu/~rik/FOA

variations without changing any code. Source code for the routines is also provided for those programmers who want to modify or extend the basic functionalities.

Exercises are collected at the end of each chapter, but they are an admittedly uneven mix. They are intended as basic review exercises; some are more challenging than others. The primary assignments for my classes are a series of machine problems: extended programming assignments that cumulatively build all the parts of a basic search engine. The details of these assignments, as well as lecture slides, test questions, and so on, are available on the ***FOA Web site***[3] to instructors who might be interested.

The first chapter of the text is designed to give any audience a broad overview of the basic questions underlying FOA and how they interact. The next three chapters cover the core issues involved in building and evaluating a generic search engine at a level appropriate to undergraduates. Chapter 5 collects several important topics that require more mathematical sophistication, and Chapters 6 and 7 consider extensions of the basic core material at a graduate level. Chapter 6 considers extensions of basic search technologies that use features of documents beyond keywords to draw more "artificially intelligent" inferences about them. Chapter 7 focuses on how one particular branch of AI, machine learning, has been used to automatically learn more about both documents and the users searching through them. Chapter 8 concludes with some looks into the most active development in FOA and a reassessment of fundamental issues that will be with us for the foreseeable future.

Acknowledgments

I had the good fortune to have David Blair at the University of Michigan (in a single lecture!) make it clear that FOA isn't just an engineering problem, but important to anyone deeply interested in language. Mike Gordon (energized by that same lecture), Manfred Kochen, Bob Lindsay, Gary and Judy Olson, Ken Winter, and Maurita Holland were all in Ann Arbor, and they taught me more than I would really appreciate until years later.

[3] http://www.cs.ucsd.edu/~rik/FOA/

Keith van Rijsbergen's unswerving confidence has made this book possible. His book is where I began and the standard I have tried to maintain. Gerry Salton and Karen Sparck Jones have been generous and patient with me as they have been to so many others in the IR community. I thank Nick Belkin, Bruce Croft, Doug Cutting, Sue Dumais, Norbert Führ, David Lewis, Jan Petersen, and Steve Robertson for uncountable interesting SIGIR dinners. I am happy to acknowledge the influence of the industrious groups around Carnegie Mellon University and Just Research, led by Tom Mitchell and Andrew McCallum, especially on Chapter 7.

A summer of exciting conversation (1987) with Ed Hutchins and Don Norman of UCSD's Cognitive Science department helped me think more broadly about "parallel distributed processing" models of cognition, involving networks of *people* rather than neurons, as parts of social systems. I have benefited from a long, productive relationship with the editors and others working at *Encyclopædia Britannica.* I am grateful to have met Mortimer Adler (once!) and especially to have worked closely with Editor-in-Chief Bob McHenry and others at *Encyclopædia Britannica* in Chicago, Chris Needham (in London), and Bob Clarke, John Dimm, John McInerney, and Harold Kester in La Jolla. Over an even longer period, Jack Conrad, Dan Dabny, Andy Desmond, Peter Jackson, and Isabelle Moulinier of West Publishing have provided my second, extended experience with the highly edited WESTLAW corpus. I enjoyed a pleasant sabbatical at the University of Wisconsin in Madison, teaching with and learning from Jude Shavlik and Mark Craven. Paul Kube is, more than anyone else I know, comfortable in both of Snow's cultures (and several others as well); he has helped me sober and balance many aspects of this manuscript. I thank Kim Itkonen for turning my words about words into a wonderful image for the cover.

Most of my own research has been done in collaboration with students. Many of my thoughts about what I had done right and wrong with AIR were shaped in conversations with Dan Rose, concerning his thesis. I am also grateful to both Dan and Susan Gruber for their help in shaping very early drafts of all chapters. Brian Bartell asked hard questions about FOA from the beginning, and I have appreciated the pleasure of his collaboration ever since. John Hatton, Amy Steier, and Fil Menczer have all helped me explore aspects of FOA as part of their own research; Thomas Kammeyer, Chris Vogt, and Bryan Tower have all helped push various aspects of the FOA code base forward. I am also grateful to Apple Computer, *Encyclopædia Britannica,* and the National Science Foundation for funding various portions of our work over the years.

Chris Rosin and Terry Jones provided useful feedback on some chapters, and Marti Hearst (University of California, Berkeley) and Paul Thompson (University of Minnesota and St. Thomas University) used early drafts of *FOA* with their classes. I am grateful to David Tranah and Shari Chappell for their rescue of *FOA* at Cambridge University Press.

Will, Lee, Cori and Julie are my nearest and dearest family. Simply completing this book (finally!) is the best apology I can offer them. Beyond that . . . "Whereof one cannot speak, one must remain silent."

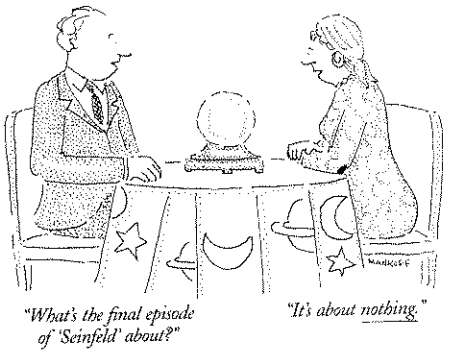
It is here where I must say that despite the best efforts of these many friends and colleagues, I know I haven't said it all, and that mistakes surely remain. I have written down those things I wish I'd known when I began my thesis, for use by students in the classes I teach. If it helps you avoid any of the mistakes it has taken me a decade to learn, it will almost have been worth it.

TERMS INTRODUCED IN THIS CHAPTER

domain of discourse
hypertext
noise words
Zipf's law

1

Overview



Finding Out About. Reproduced by permission of *The New Yorker*[*]

1.1 Finding Out About – A Cognitive Activity

We are all forced to make decisions regularly, sometimes on the spur of the moment. But the rest of the time we have enough warning that it is possible to collect our thoughts and do some research that makes our

[*] Robert Mankoff, © *The New Yorker*, 26 January 1998.

decision as sound as it can be. This book is a closer look at the process of **finding out about** (FOA), research activities that allow a decision-maker to draw on others' knowledge. It is written from a technical perspective, in terms of computational tools that speed the FOA activity in the modern era of the distributed networks of knowledge collectively known as the World Wide Web (WWW). It shows you how to build many of the tools that are useful for searching collections of text and other media. The primary argument advanced is that progress requires that we appreciate the *cognitive* foundation we bring to this task as academics, as language users, and even as adaptive organisms.

As organisms, we have evolved a wide range of strategies for seeking useful information about our environment. We use the term "cognitive" to highlight the use of *internal representations* that help even the simplest organisms perceive and respond to their world; as the organisms get less simple, their cognitive structures increase in complexity. Whether done by simple or complex organisms, however, the process of *finding out about* is a very active one – making initial guesses about good paths, using complex sets of features to decide if we seem to be on the right path, and proceeding forward.

As humans, we are especially expert at searching through one of the most complex environments of all: *language*. Its system of linguistic features is not derived from the natural world, at least not directly. It is a constructed, cultural system that has worked well since (by definition!) prehistoric times. In part, languages remain useful because they are capable of change when necessary. New features and new objects are noticed, and it becomes necessary for us to express new things about them, to form our reactions to them, and to express these reactions to one another.

Our first experience of language, as children and as a species, was oral – we spoke and listened. As children we learn *Sprachspiele* (word or language games) [Wittgenstein, 1953] – how to use language to get what we want. A baby saying "Juice!" is using the exclamation as a *tool* to make adults move; that's what a word means. Such a functional notion of language, in terms of the jobs it accomplishes, will prove central to our conception of what keywords in documents and queries mean as part of the FOA task.

Beyond the oral uses of language, as a species we have also learned the advantages of *writing down* important facts we might otherwise forget.

Writing down a list of things to do, which we might forget tomorrow, extends our limited memory. Some of these advantages accrue to even a single individual: We use language personally, to organize our thoughts and to conceive strategies.

Even more important, we use writing to say things to others. Writing down important, memorable facts in a consistent, **conventional** manner, so that others can understand what we mean and vice versa, further amplifies the linguistic advantage. As a society, we value reading and writing skills because they let us interpret shared symbols and coordinate our actions. In advanced cultures' scholarship, entire curricula can be defined in terms of what Robert McHenry (Editor-in-Chief of *Encyclopædia Britannica*) calls ***"Knowing How to Know."***[1]

It is easiest to think of the organism's or human's search as being for a valuable object, sweet pieces of fruit in the jungle, or (in modern times) a grocer that sells them. But as language has played an increasingly important role in our society, searching for valuable written passages becomes an end unto itself. Especially as members of the academic community, we are likely to go to libraries seeking others' writings as part of our search. Here we find rows upon rows of books, each full of facts the author thought important, and endorsed by a librarian who has selected it. The authors are typically people far from our own time and place, using language similar but not identical to our own.

Of course the library contains many such books on many, many topics. We must Find Out About a topic of special interest, looking only for those things that are **relevant** to our search. This basic skill is a fundamental part of an academic's job:

- We look for references in order to write a term paper.
- We read a textbook, looking for help in answering an exercise.
- We comb through scientific journals to see if a question has already been answered.

We know that if we find the right reference, the right paper, the right paragraph, our job will be made much easier. Language has become not only the means of our search, but its object as well.

[1] www.justanother.com/howtoknow

Today we can also search the **World Wide Web** (WWW) for others' opinions of music, movies, or software. Of course these examples are much less of an "academic exercise"; Finding Out About such information commodities, and doing it consistently and well, is a skill on which the modern information society places high value indeed. But while the infrastructure forming the modern WWW is quite recent, the promise offered by truly connecting all the world's knowledge has been anticipated for some time, for example, by H. G. Wells [Wells, 1938].

Many of the FOA searching techniques we will discuss in this text have been designed to operate on vast collections of apparently "dead" linguistic objects: files full of old email messages, CD-ROMs full of manuals or literature, Web servers full of technical reports, and so on. But at their core, each of these collections is evidence of real, vital attempts to communicate. Typically an **author** (explicitly or implicitly) anticipates the interests of some imagined **audience** and produces text that is a balance between what the author wants to say and what he or she thinks the audience wants to hear. A textual **corpus** will contain many such documents, written by many different authors, in many styles and for many different purposes. A person searching through such a corpus comes with his or her own purposes and may well use language in a different way from any of the authors. But each individual linguistic expression – the authors' attempts to write, the searchers' attempts to express their questions and then read the authors' documents – must be appreciated for the **word games** [Wittgenstein, 1953] that they are. FOA is centrally concerned with meaning: the semantics of the words, sentences, questions, and documents involved. We cannot tell if a document is about a topic unless we understand (at least something of) the semantics of the document and the topic. This is the notion of about-ness most typical within the tradition of library science [Hutchins, 1978].

This means that our attempts to engineer good technical solutions must be informed by, and can contribute to, a broader philosophy of language. For example, it will turn out that FOA's concern with the semantics of entire documents is well complemented by techniques from computational linguistics, which have tended to focus on syntactic analysis of individual sentences. But even more exciting is the fact that the recent

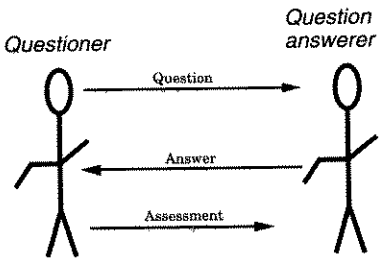


FIGURE 1.1 The FOA Conversation Loop.

availability of new types of **electronic artifacts** – from email messages and WWW corpora to the browsing behaviors of millions of users all trying to FOA – brings an *empirical* grounding for new theories of language that may well be revolutionary.

At its core, the FOA process of browsing readers can be imagined to involve three phases:

1. asking a question;
2. constructing an answer; and
3. assessing the answer.

This conversational loop is sketched in Figure 1.1.

Step 1. Asking a Question

Meta-cognition about ignorance

The first step is initiated by people who (anticipating our interest in building a search engine) we'll call **users**, and their questions. We don't know a lot about these people, but we do know they are in a particular frame of mind, a special cognitive state; they may be aware† of a specific gap in their knowledge (or they be only vaguely puzzled), and they're motivated to fill it. They want to FOA . . . some topic.

Supposing for a moment that we were there to ask, the users may not even be able to characterize the topic, that is, to articulate their knowledge gap. More precisely, they may not be able to fully define characteristics of the "answer" they seek. A paradoxical feature of the FOA problem is

that if users knew their question, precisely, they might not even need the search engine we are designing: Forming a clearly posed question is often the hardest part of answering it! In any case, we'll call this somewhat befuddled but not uncommon cognitive state the users' **information need.**

While a bit confused about their particular question, the users are not without resources. First, they can typically take their ill-defined, *internal* cognitive state and turn it into an *external* expression of their question, in some language. We'll call their expression the **query**, and the language in which it is constructed the **query language.**

Step 2. Constructing an Answer

So much for the source of the question; whence the answer? If the question is being asked of a person, we must worry about equally complex characteristics of the *answerer's* cognitive state:

- Can they translate the user's ill-formed question into a better one?
- Do they know the answer themselves?
- Are they able to verbalize this answer?
- Can they give the answer in terms the user will understand?
- Can they provide the necessary background knowledge for the user to understand the answer itself?

We will refer to the question-answerer as the **search engine**, a computer program that algorithmically performs this task. Immediately each of the concerns (just listed) regarding the *human* answerer's cognitive state translates into extremely ambitious demands we might make of our *computer* system.

Throughout most of this book, we will avoid such ambitious issues and instead consider a very restricted form of the FOA problem: We will assume that the search engine has available to it only a set of preexisting, "canned" passages of text and that its response is limited to identifying one or more of these passages and presenting them to the users; see Figure 1.2. We will call each of these passages a **document** and the entire set of documents the corpus. Especially when the corpus is very large (e.g., assume it contains millions or even billions of documents), selecting a very small set (say 10 to 20) of these as potentially good answers to

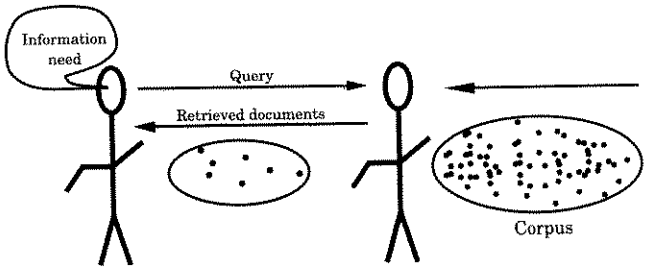


FIGURE 1.2 Retrieval of Documents in Response to a Query

be **retrieved** will prove sufficiently difficult (and practically important) that we will focus on it for the first few chapters of this book. In the final chapters however, we will consider how this basic functionality can be extended towards tools for "Searching for an education" (cf. Section 8.3.9).

Step 3. Assessing the Answer

Imagine a special instance of the FOA problem: You are the user, waiting in line to ask a question of a professor. You're confused about a topic that is sure to be on the final exam. When you finally get your chance to ask your question, we'll assume that the professor does nothing but select the three or four preformed pearls of wisdom he or she thinks come closest to your need, delivers these "documents," and sends you on your way. "But wait!" you want to say. "That isn't what I meant." Or, "Let me ask it another way." Or, "That helps, but I still have this problem."

The third and equally important phase of the FOA process "closes the loop" between asker and answerer, whereby the user (asker) provides an assessment of how relevant they find the answer provided. If after your first question and the professor's initial answer you are summarily ushered out of the office, you have a perfect right to be angry because the FOA process has been violated. FOA is a *dialog* between asker and answerer; it does not end with the search engine's first delivery of an answer. This initial exchange is only the first iteration of an ongoing conversation by which asker and answerer mutually negotiate a satisfactory exchange. In the process, the asker may *recognize* elements of the answer he or she

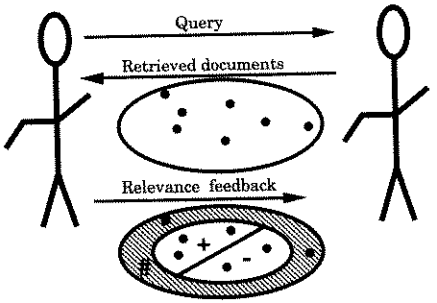


FIGURE 1.3 Assessment of the Retrieval

seeks and be able to reexpress the information need in terms of threads taken from previous answers.

Because the question-answerer has been restricted to a simple set of documents, the asker's **relevance feedback** must be similarly constrained; for each of the documents retrieved by the search engine, the asker reacts by saying whether or not the document is relevant. Returning to the student/professor scenario, we can imagine this as the student saying "Thanks, that helps" after those pearls that do and remaining silent or saying, "Huh?" or "What does that have to do with anything?!" or "No, that's not what I meant!" otherwise. More precisely, relevance feedback gives askers the opportunity to provide more information with their reaction to each retrieved document – whether it is relevant (⊕), irrelevant (⊖), or neutral (#). This is shown as a Venn diagram–like labeling of the set of retrieved documents in Figure 1.3. We'll worry about just how to solicit and make use of relevance feedback judgments in Chapter 4.†

What FOA data can we observe?

1.1.1 Working within the IR Tradition

If it seems to you that the last section has sidestepped many of the most difficult issues underlying FOA, you're right! Later chapters will return to redress some of these omissions, but the immediate goal of Chapters 2 to 4 is to "operationalize" FOA to resemble a well-studied problem within computer science, typically referred to as **information retrieval** (IR). IR is a field that has existed since computers were first

used to count words [Belkin and Croft, 1987]. Even earlier, the related discipline of library science had developed many automated techniques for efficiently storing, cataloging, and retrieving *physical* materials so that browsing patrons could find them; many of these methods can be applied to the digital documents held within computers. IR has also borrowed heavily from the field of linguistics, especially computational linguistics.

Other places to FOA IR

The primary journals in the field and most important conferences† in IR have continued to publish and meet since the 1960s, but the field has taken on new momentum within the last decade. Computers capable of searching and retrieving from the entire biomedical literature, across an entire nation's judicial system, or from all of the major newspaper and magazine articles, have created new markets among doctors, lawyers, journalists, students, everyone! And of course, the Internet, within just a few years, has generated many, many other examples of textual collections and people interested in searching through them.

The long tradition of IR is therefore the primary perspective from which we will approach FOA. Of course, every tradition brings with it tacit assumptions and preconceived notions that can hinder progress. In some ways, an elementary school student using the Internet to FOA class materials is related to the original problem considered by library science and IR, but in many other ways it couldn't be more different (cf. Section 8.1). In this text, "FOA" will be used to refer to the broadest characterization of the cognitive process and "IR" to this subdiscipline of computer science and its traditional techniques. When we talk of the "search engine," this is not meant to refer to any particular implementation, but to an idealized system most typical of the many different generations and varieties of actual search engines now in use. If you are using this text as part of a course, you may build one simple example of a search engine.

Using Figure 1.4 as a guide, we'll return to each of the three phases and be a bit more specific about each component of our search engine. Here, finally, the human question-answerer has been replaced by an algorithm, the search engine, that will attempt to accomplish the same purpose. This figure also makes clear that the fundamental operation performed by a search engine is a *match*, between descriptive features mentioned by users in their queries and documents sharing those features. By far the most important kind of features are keywords.

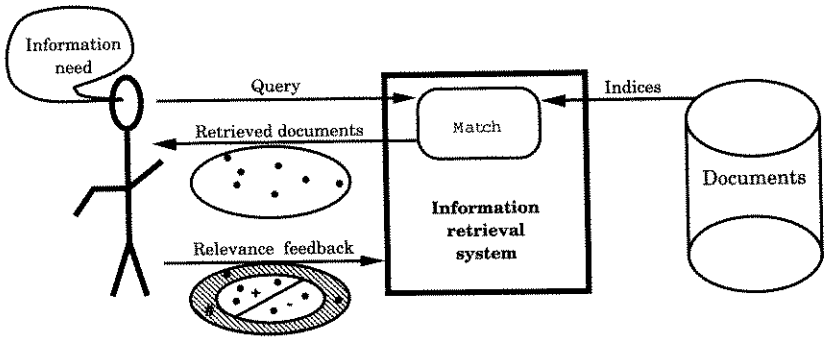


FIGURE 1.4 Schematic of Search Engine

1.2 Keywords

Keywords are linguistic atoms – typically words, pieces of words, or phrases – used to characterize the subject or content of a document. They are pivotal because they must bridge the gap between the users' characterization of information need (i.e., their queries) and the characterization of the documents' topical focus against which these will be matched. We could therefore begin to describe them from either perspective: how they are used by users, or how they become associated with documents. We will begin with the former.

1.2.1 Elements of the Query Language

If the query comes from a student during office hours or from a patron at a reference librarian's desk, the query language they'll use to frame their question is entirely **natural**, that most expressive "mother tongue" familiar to both question-asker and -answerer. But for the software search engines we will consider, we must assume a much more constrained "artificial" query language. Like other languages, ours will have both a meaningful **vocabulary** – the set of important keywords any user is allowed to mention in any queries – and a *syntax* that allows us to construct more elaborate query structures.

1.2.2 Topical Scope

The first constraint we can apply to the set of keywords we will allow in our vocabulary is to define a domain of discourse – the subject area within which each and every user of our search engine is assumed to be searching. While we might imagine building a truly encyclopedic reference work, one capable of answering questions about any topic whatsoever, it is much more common to build a search engine with more limited goals, capable of answering questions about some particular subject. We will choose the simpler path (it will prove enough of a challenge!) and focus on a particular topic. To be concrete, throughout this text we will assume that the domain of discourse is ARTIFICIAL INTELLIGENCE (AI). Briefly, AI can be defined as a subdiscipline of computer science, especially concerned with algorithms that mimic inferences which, had they been made by a human, would be considered "intelligent." It typically includes such topics as KNOWLEDGE REPRESENTATION, MACHINE LEARNING, and ROBOTICS.

Thus COMPUTER SCIENCE is a **broader term** than ARTIFICIAL INTELLIGENCE. This **hypernym** relationship between the two phrases is something we will return to later (cf. Section 6.3). For example, our task becomes more difficult if we assume that the corpus of documents contains material on the broader topic of COMPUTER SCIENCE, rather than just (!) ARTIFICIAL INTELLIGENCE. Conversely, the topics KNOWLEDGE REPRESENTATION, MACHINE LEARNING, and ROBOTICS are all **narrower terms**, and our task would, *caeteris paribus*,* be made easier if we only had to help users FOA one of them.

Constraining the vocabulary so that it is **exhaustive** enough that any imaginable and relevant topic is expressible within the language, while remaining **specific** enough that any particular subjects a user is likely to investigate can be distinguished from others, will become a central goal of our design. ROBOTICS, for example, would seem a descriptive keyword because it identifies a relatively small subarea of ARTIFICIAL INTELLIGENCE. COMPUTER SCIENCE would be silly as a keyword (for this corpus), because we are assuming it would apply to every document and hence does nothing to discriminate them – it is too exhaustive. At the

* (Assuming) all other things are equal.

other extreme, ROBOTIC VACUUM CLEANERS FOR 747 AIRLINERS is almost certainly too specific.

The **vocabulary size** – the total number of keywords – depends on many factors, including the scope of the domain of discourse. A typical language user has a reading vocabulary of approximately 50,000 words. Web search engines and large test corpora formed from the union of many document types may require vocabularies ten times this size. It is unlikely that such a large lexicon of keywords would be required for restricted corpora, but it is also true that even a narrow field can develop an extensive, specialized **jargon** or **terms of art**. In practice, search engines typically have difficulty reducing the number of usable keywords to much below 10,000.

1.2.3 Document Descriptors

We've introduced keywords as features mentioned by users as part of their queries, but the other face of keywords is as descriptive features of documents. That is, we might naturally say that a document is about ROBOTICS. Users mentioning ROBOTICS in their query should expect to get those documents that are about this topic. Keywords must therefore also function as the *documents'* description language. The same vocabulary of words used in queries must be used to describe the topical content of each and every document. Keywords become our characterization of what each document is about. **Indexing** is the process of associating one or more keywords with each document.

The vocabulary used can either be **controlled** or **uncontrolled** (a.k.a. **closed vocabularies** or **open vocabularies**). Suppose we decide to have all the documents in our corpus manually indexed by their authors; this is quite common in many conference proceedings, for example. If we provide a list of potential keywords and tell authors they must restrict their choices to terms on this list, we are using a controlled indexing vocabulary. On the other hand, if we allow the authors to assign any terms they choose, the resulting index has an uncontrolled vocabulary [Svenonius, 1986].

To get a feel for the indexing process, imagine that you are given a piece of text and must come up with a set of keywords that describe what the document is about. Let's make the exercise more concrete. You are the author of a report entitled USING A NEURAL NETWORK FOR PREDICTION, and you are submitting it to a journal. One of the things this particular journal requires is that the author provide up to six keywords under which this article will be indexed. If you are sending it to the *Communications of the ACM*, you might pick a set of keywords that identify, to the audience of computer scientists you think read this publication, connections between this new work and prior work in related areas: NONLINEAR REGRESSION; TIME SERIES PREDICTION.

But now imagine that you've decided to submit the *exact same paper* to *Byte* magazine, and you must again pick keywords that have meaning to this audience. You might choose: NEURAL NETWORKS; STOCK MARKET ANALYSIS.

What is the **context** in which these keywords are going to be interpreted? Who's the audience? Who's going to understand what these keywords mean? Anticipating the FOA activity in which these keywords will function, we know that the real issue to be solved is not only to describe this one document, but to *distinguish* it from the millions of others in the same corpus. How are the keywords chosen going to be used to distinguish your document from the others?

It is often easiest to imagine keywords as independent features of each document. In fact, however, keywords are best viewed as a *relation* between a document and its prospective readers, sensitive to both characteristics of the users' queries and other documents in the same corpus. In other words, the keywords you pick for *Byte* should be different from those you pick for *Communications of the ACM*, and for deeper reasons than what we might cynically consider "spin control."

1.3 Query Syntax

Keywords therefore have a special status in IR and as part of the FOA process. Not only must they be exhaustive enough to capture the entire topical scope reflected by the corpora's domain of discourse, but they must also be expressive enough to characterize any information needs the users might have.

Of course we need not restrict our users to only one of these keywords. It seems quite natural for queries to be composed of two or three, perhaps even dozens, of keywords. Recent empirical evidence suggests that many typical queries have only two or three keywords

(cf. Section 8.1), but even this number provides a great combinatorial extension to the basic vocabulary of single keywords. Other applications, for example, using a document itself as a query (i.e., using it as an example: "Give me more like this"), can generate queries with hundreds of keywords. Regardless of size, queries defined only as sets of keywords will be called **simple queries**. Many Web search engines support only simple queries. Often, however, the search engines also provide more advanced interfaces, including **operators** in the query language. Perhaps, because you have previously been warped by an exposure to computer science:), you think that sets of keywords might be especially useful if joined by Boolean operators. For example, if we have one set of documents about NEURAL NETWORKS and another set of documents about SPEECH RECOGNITION, we can expect the query: NEURAL NETWORKS AND SPEECH RECOGNITION to correspond to the intersection of these two sets, while NEURAL NETWORKS OR SPEECH RECOGNITION would correspond to their union.

The Boolean NOT operator is a bit more of a problem. If users say they want things that are *not* about NEURAL NETWORKS, they are in fact referring to the vast majority of the corpus. That is, NOT is more appropriately considered a binary, subtraction operator. To make this distinction explicit we will call it BUT_NOT.

There are other syntactic operators that are often included in a search engine's query language, but discussion of these will be put off until later. Even with these simple Boolean connectives and a keyword vocabulary of reasonable size, users can construct a vast number of potential queries when attempting to express their information need.

1.3.1 Query Sessions

As we consider the specific features of each query, it is important to remember the role these short expressions play in the larger FOA process. Queries are generated as an attempt by users to express their information need. As with any linguistic expression, conveying a thought you have can be difficult, and this is likely to be especially true of the muddled cognitive state of our FOA searcher. Users who are familiar with the special syntactic features of a query language may be able to express their need more easily, but others for whom this unnatural syntax is new or difficult will have additional difficulties.†

"Typical" users have changed

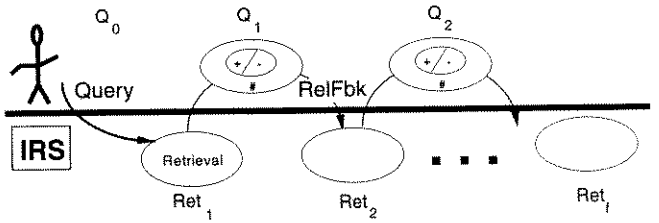


FIGURE 1.5 A Query Session

As with many of the idealizing assumptions we are at least temporarily making, it is often simpler to think about only one iteration of the three-step query/retrieve/assess FOA process at a time. In most realistic situations we can expect that single queries will not occur in isolation but as part of an iteration of the FOA process. An initial query begins the dialog; the search engine's response provides clues to the user about directions to pursue next; these are expressed as another query. An abstract view of this sequence is presented in Figure 1.5. Note especially the concatenation of a series of basic FOA three-step iterations. Data are produced by the user, then by the search engine, and then by the user; this constructs a very natural alternation of user–search engine exchanges. Users' assessments can also function as their next query statement. This can be achieved simply if we have some method for automatically constructing a query from relevance feedback. For example, if users click on documents they like, the search engine can, by itself, form a new query that focuses on those keywords that are especially associated with these documents.

There are many such techniques for using relevance feedback from a single query/retrieval, and there are many more things we can learn from the entire query session. The full query session provides more complete evidence about the users' information need than we can gain from any one query. In fact, as will be discussed extensively in Chapter 7, there exist algorithmic means by which the search engine itself might "learn" from such evidence. Learning methods might even be expected to make **transitive** leaps, from the users' initial expressions of their information needs to the final documents that satisfied them.† (Of course, this transitive leap is only warranted if we are certain that users ended the session

Transitivity

satisfied and aren't just quitting in frustration!) For all these reasons, we must try to identify a query session's boundaries, that is, when one focused search session ends and the next session, involving the same user searching on a different topic, begins.

1.4 Documents

When "documents" were first introduced as part of the FOA process, it was as one of the set of potential, predefined answers to users' queries. Here we will ground this abstract view in practical terms that can be readily applied, for example, to the searches that are now common on the Web. Our goal will be to balance this practical description of how search engines work today with the abstract FOA view that goes beyond current practices to other kinds of searches still to come.

A useful working definition is that a document is a *passage of free text*. It is composed of text, strings of characters from an alphabet. We'll typically make the (English) assumption that uses the Roman alphabet, Arabic numerals, and standard punctuation. Complications like font style (italics, bold) and non-Roman **marked alphabets** that add characters like ä, Ç, Ñ, and æ; and the iconic characters of Asian languages require even more thought.

By "free" text we mean it is in **natural language**, the sort native readers and writers use easily. Good examples of free text might be a newspaper article, a journal paper, or a dictionary definition. Typically the text will be grammatically well-formed language, in part because this is *written* language, not oral. People are more careful when constructing written artifacts that last beyond the moment. Informal texts like email messages, on the other hand, help to point to ways that some texts can retain the spontaneity of oral communication, for better and worse [Ong, 1982].

Finally, we will be interested in **passages** of such text, of arbitrary size. The newspaper example makes us imagine documents of a few thousand words, but journal articles make us think of samples ten times larger, and email messages make us think of something only a tenth that size. We can even think of an entire book as a single document. All such passages satisfy our basic definition; they might be appropriate answers to a search about some topic.

The length of the documents will prove to be a critical issue in FOA search engine design, especially when some corpus contains documents of widely varying lengths. This is because longer documents can discuss more topics, so they are capable of being about more. Longer documents are more likely to be associated with more keywords, and hence they are more likely to be retrieved (cf. Section 3.4.2).

One possible response is to make a simple but very consequential assumption.

ASSUMPTION 1 *All documents have equal about-ness.*

In other words, if we ask the (a priori) probability of any document in the corpus being considered relevant, we will assume that all are equiprobable. This would lead us to *normalize* documents' indices in some way to compensate for differing lengths. The normalization procedure is a matter of considerable debate; we will return to consider it in depth later (cf. Section 3.4.2).

For now, we will take a different tack toward the issue of document length, as captured by an alternative pair of assumptions.

ASSUMPTION 2 *The smallest unit of text with appreciable about-ness is the paragraph.*

ASSUMPTION 3 *All manner of longer documents are constructed out of basic paragraph atoms.*

The first piece of this argument is that the smallest sample of text that can reasonably be expected to satisfy an FOA request is a paragraph. The claim is that a word, even a sentence, does not by itself provide enough *context* for any question to be answered or "found out about." But if the paragraph has been well constructed, as defined by conventional rules of composition, it should answer many such questions. And unless the text comes from James Joyce, Proust, or Jorge Luis Borges, we can expect paragraphs to occupy about half an average screen page – nicely viewable chunks.

Assumption 3 alludes to the range of structural relationships by which the atomic paragraphs can typically be strung together to form longer passages. First and foremost is simple sequential flow, the order

in which an author expects the paragraphs to be read. The sequential nature of traditional printed media, from the first papyrus scrolls to modern books and periodicals, has meant that a sequential ordering over paragraphs has been dominant. It may even be that the modern human is especially capable of understanding **rhetoric** of this form (cf. Section 6.2.3).

In any case, a sequential ordering of paragraphs is just one possible way they might be related. Other common relationships include:

- a *hierarchical* structure composing paragraphs into subsections, sections, and chapters;
- *footnotes*, embellishing the primary theme;
- *bibliographic citations* to other, previous publications;
- *references* to other sections of the same document; and
- *pedagogical prerequisite* relationships ensuring that conceptual foundations are established prior to subsequent discussion.

Of course each of these relationships has grown up within the tradition of printed publication. Special typographical conventions (boldface, italics, sub- and superscripting, margins, rules) have arisen to represent them and distinguish them from sequential flow.

But new, electronic media now available to readers (and becoming available to authors) need not follow the same strictly linear flow. The new capabilities and problems of traversing text in nonlinear ways – hypertext – have been discussed by some visionaries [Bush, 1945; Nelson, 1987] for decades. This new technology certainly permits us to make some traversals more easily (e.g., jumping to a cited reference with the click of a button rather than a trip to the library), but this same ease may make it more difficult for an author to present a cogent argument.

For now we will not worry about how arguments can be formed with nonlinear hypermedia. Assumptions 2 and 3 simply allow us to infer Assumption 1: If all the documents are paragraphs, we can expect them to have virtually uniform about-ness. These are also simplifying assumptions, however. In an important sense, a scientific paper's abstract is about the same content as the rest of the paper, and a newspaper article's first paragraph attempts to summarize the details of the following story. These issues of a text's **level of treatment** will be discussed later.

1.4.1 Structured Aspects of Documents

In addition to their free text, many documents will carry **meta-data** that gives some facts about the document. We may have **publication information**, for example, that this document appeared in this journal, in this issue, on this page. We are likely to know the author(s) of the document. Queries will often refer to aspects of both free text and meta-data.

QUERY 1 *I'm interested in documents about Fools' Gold that have been published in children's magazines in the last five years.*

The first portion of this query depends on the same about-ness relation that is at the core of our FOA process. But the last two criteria, concerning publication type and date, seem to be just the sort of query against structured attributes that database systems perform very successfully. In most real-life applications a hybrid of database and IR technologies will be necessary. (We distinguish between these techniques in Section 1.6.)

The most interesting examples concern characteristics that do not clearly fall into either IR or database categories. For example, can you define precisely what you mean by a "children's magazine" in terms of unambiguous attributes on which a database would depend? Consider another query.

QUERY 2 *What sort of work has K. E. Smith done on metabolic proteins affecting neurogenesis?*

Finding an exact match for the string `K. E. Smith` in the `AUTHORS` attribute is straightforward. But the conventions in much of medical and biological publication (as well as in some areas of physics) sometimes lead to dozens of authors on papers, from the director of the institute through all of the laboratory assistants. Although `K. E. Smith` might well fulfill the syntactic requirements of authorship on a particular paper, users searching for "the work of" this person might well have a more narrowly defined *semantic* relationship in mind.

1.4.2 Corpora

We have focused on individual documents, but of course the FOA problem would not interest us except that we are typically faced with a corpus of *millions* of such documents, and we are interested in finding only the handful that are of interest. The actual number of documents and their cumulative size will matter a great deal, as some of our IR methods have time or space complexities that make them viable only within certain parameters. To pick a simple example, if you are trying to find a newspaper article (you read it a few days ago) for a friend, exhaustively searching through all the pages is probably quite effective if you know it was in Friday's paper, but not if you need to search through an entire month's recycling pile! Similarly, a standard utility like the Unix grep command can be a practical alternative if the corpus is small and the queries simple.

1.4.3 Document Proxies

Do you remember the library's original card catalogs, those wooden, beautifully constructed cabinets full of rows and rows of drawers, each full of carefully typed index cards? The card catalog contained **proxies** – abridged representations of documents, acting as their surrogate – for the books it indexed. No one expected the full text of the books to actually be found in the drawers.

Computerized card catalogs are only capable of supporting a similar function. They do allow more extensive indexing and efficient retrieval, from terminals that might be accessed far from the library building. At the heart of this system is a text search engine capable of matching features of a query against book titles.† Just like with the original index cards, however, retrieval is limited to some proxy of the indexed work, a bibliographic citation, or perhaps even an abstract. The text of ultimate interest – in a book, magazine, or journal – remains physically quite distinct from the search engine used to find it.

Card catalogs were the first search engines

As computer storage capacities and network communication rates have exploded, it has become increasingly common to find retrieval systems capable of presenting the full text of retrieved items. In the modern context, proxies extend beyond the bibliographic citation information and subject headings we associate with card catalogs and include a

document's title, an article's abstract, a judicial opinion's headnote, or a book's table of contents.

The distinction between the search engine retrieving documents and retrieving proxies remains important, however, for at least two reasons. First, the radically changing technical capabilities of libraries (and computers and networks more generally) can create conceptual confusion about just what the search engine is doing. While it has been possible for a decade or more to get the full text of published journal articles through commercial systems such as DIALOG and Lexis/Nexis, free access to these through your public library would have been almost unheard of until quite recently. In fact, most libraries did not even try to index individual articles in their periodical collections. Changing technical capacities, changes in the application of intellectual property laws, changes in the library's role, and resulting changes in the publishing industry are radically altering the traditional balance. Even when all new publications are easily available electronically, the issue of *retrospectively* capturing previously published books and journals remains unresolved.

Looking far into the future and assuming no technical, economic, or legal barriers to a complete rendering of any document in our corpus, there is still an important reason to consider document proxies. Recall that FOA is a *process* we are attempting to support and that retrieving sets of documents to show users is a step we expect to repeat many times. Proxies are abridged versions of the documents that are easier for browsing users to quickly scan and react to (i.e., provide relevance feedback) than if they had to read the entire document. If a document's title is accurate (if its abstract is well written, if its bibliographic citation is complete), this proxy may provide enough information for users to decide if it seems relevant.†

A misleading title, or did the document teach you something?!

1.4.4 Genre

A more subtle characteristic of documents that may need to concern us is their **genre** – the voice or style in which a document is written. You would, um, like, be pretty darn surprised to find stuff like this in a textbook, but not if it came to you over the phone. The genre of email seems to be settling somewhere between typical printed media and spoken conversation, with special markings of sarcasm:) and expletives #!?% common. Newspaper journalists are carefully trained to produce articles

consistent with what newspaper readers expect, and their editors are paid to ensure that these stories maintain a consistent voice. Scientific journal articles are written to be understood by peers in the same field, according to standards that pertain to that community [Bayerman, 1988]. An important component of this audience focus is the **vocabulary choice** an author makes (cf. Section 8.2.1); stylistic variations and document structure may also differ. In a field like psychology, for example, it would be difficult to get a paper accepted in some journals if it is not subdivided into sections like Hypothesis, Methodology, and Subjects. Legal briefs are also written in highly conventionalized forms [Havard Law Review Association, 1995], and legislation is drafted to satisfy political realities [Allen, 1980; Goodrich, 1987; Levi, 1982; Nerhot, 1991].

In part, these variations in genre are difficult to detect because they remain consistent within any single corpus. That is, the typical email message would jump out at you as out of place if it appeared in your newspaper, but probably not if it were on the Letters to the Editor page. These examples highlight how much *context* about the corpus we bring with us whenever we read a particular document. They also foreshadow problems Web searchers are just beginning to appreciate, as WWW search engines include every document to which they can crawl, intermixing their very different contexts and writing styles. Without the orienting features of the newspaper's masthead, the "Letters to the Editor" rubric, or the purposeful selection of a tool that scans only Usenet news, the browsing users' abilities to understand an arbitrary document is diminished. Individual textual passages have been stripped of much of the context that made them sensible. As more and more of us generate *content* – in new hypermedia forms as well as traditional publications – that more and more of us retrieve, the range of genres we will experience can only increase, and our methods for FOA must help to represent not just the document but contextual information as well.

1.4.5 Beyond Text

Our definition of "documents" has hewn closely to the printed forms that still dominate the FOA retrievals most people now do. But print media are not the only form of answer we might reasonably seek, and we must ensure that our methods generalize to the other media that are

increasingly part of the Net. Sound, images, movies, maps, and more are all appearing as part of the WWW, and they are typically intermixed with textual material. We need to be able to search all of these.

One reason for casting the central problem of this text as "finding out about" is that many aspects of multimedia retrieval remain the same from this perspective. We still have users, who have information needs. We can still reasonably use the term "document" to include any potential answer to users' queries, but now we expand this term to include whatever media are available. Most centrally, we must still characterize what each document is about in order to match it to these queries, and users can still assess how well the search engine has done.

At the same time, many parts of the FOA problem change as we move away from textual documents to other media. Most important is the increased difficulty of algorithmically extracting clues related to the documents' semantic content from their syntactic features. The primary source of semantic evidence used within text-based IR is the relative frequencies of keywords in document corpora, and a major portion of this text will show that this is a powerful set of clues indeed. We will also discuss the role other syntactic clues (e.g., bibliographic links) associated with texts can play in understanding what they are about. As we move to other media, the important question becomes what consistent features these new media have that we can also process to reliably infer semantic content. For example, what can we know about an image from the distribution of its pixel values? Do all SUNSETS share a brightness profile (dark below a horizontal line, symmetrically bright above it) that is reliable enough that this clue can be exploited to identify just these scenes?[†] If so, can this mode of analysis be generalized sufficiently to allow retrieval of images based on more typical descriptors such as CHILDREN FEEDING ANIMALS?

Signature of human culture?!

Even if we imagine that certain obvious, superficial aspects of some images may be extracted, our hopes must not blind us to the rich vocabulary that many images use every day. Consider a query like FIDELITY AS A POLITICAL ISSUE and consider Figure 1.6. Would any reasonable person claim that they could provide an *exhaustive* list of all the things these pictures "say"? Did you include the set of Hillary's jaw? The angle of Bill's gaze? The attitudes about divorce prevalent when the Doles' picture was taken and now? The tacit commentary by the editors of

FIGURE 1.6 Finding Out About POLITICAL FIDELITY*
Reproduced by permission of *The New York Times*

The New York Times produced by the juxtaposition of these two photos? Note also that this picture (and its selection for use in this text!) occurred years before anyone had even heard of Monica Lewinsky![†]

MONICA the meme

Figure 1.7 gives a second example. This is a photograph of a locking display case, containing a concert performance schedule. Pasted over the glass of the case is a sign, saying: "IGNORE THIS CALENDAR: THESE DATES ARE 3 YEARS OLD." But the photo also reveals a number of more subtle clues – that the key to the case has been lost (for three years!), that some frustrated teacher finally got tired of dealing with confused parents, that none of the school's administrators can think of a more imaginative solution.

These examples may seem far-fetched. But those of you old enough to remember the Cold War may also remember that there was an entire job category known as "Kremlinologist": someone expert at divining various power shifts among the Politburo based on evidence such as where various participants were placed within group photos! The conventional wisdom is that "a picture is worth a thousand words," and although some images may not require much explanation, others speak volumes. As we move from still images to *movies*, entirely new channels

* *The New York Times*, 15 Sept. 1996, Week in Review, p. 1.

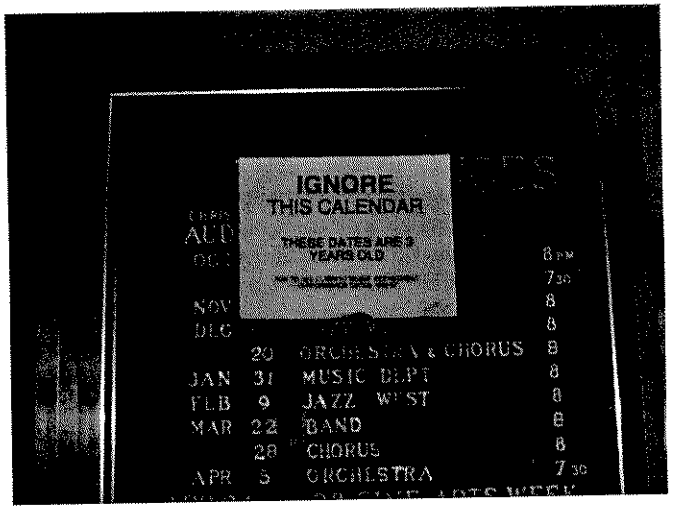


FIGURE 1.7 Obsolete Concert Schedule

for meaning – conveyed with the camera's attentional focus, soundtrack, etc. – are available to a skilled director. Music itself has an equally rich but distinct vocabulary. The ability to easily record and transmit digital **spoken documents** (speech) makes this form of audio especially worthy of analysis [Sparck Jones et al., 1996].

As with text, music, film, and motion pictures all predate their representations on computers. The convenience and availability of all these electronic media make it more possible and even more important to analyze them.

Once again, text is an excellent place to begin. **Semiotics** is one label for the subfield of linguistics concerned with words as *symbols*, as conveyors of meaning. Words in a language represent a particularly coherent system of symbol use, but so do the symbols used by photo journalists, painters, and movie directors. The meaning of these symbols changes with time; recall the pictures of the Clintons and Doles, their interpretation at the time of publication, and their interpretation now. What these pictures mean is different if we ask about the original context of 1996 and its meaning now. And again, complex, shifting meanings are

typical not only of images but of documents as well: Watson and Crick's publication of the DNA code in *Nature* in 1953 [Watson and Crick, 1953] was important even then, but what that paper means now could not have been anticipated.

Yet the prospects for associating contentful descriptors with images and even richer media are not quite as bleak as they might seem. In many important cases (e.g., the archives of news photos maintained by magazines and newspapers), images are accompanied by **captions**, and video streams with **transcripts**. This additional *manually constructed textual* data means that techniques for inferring semantic content directly from images can piggyback on top of text-based IR techniques. In conjunction with the machine learning techniques we will discuss (cf. Chapter 7), statistically reliable associations found in captioned image and video corpora can be extrapolated to situations where we have images without captions and video without transcripts.

In the interim, we will return to the narrower, text-only notion of a document with which we began and consider FOA solutions for this simpler (!) case.

1.5 Indexing

Indexing is the process by which a vocabulary of keywords is assigned to all documents of a corpus. Mathematically, an index is a *relation* mapping each document to the set of keywords that it is about:

$$Index : doc_i \xrightarrow{about} \{kw_j\}$$

The inverse mapping captures, for each keyword, the documents it **describes**:

$$Index^{-1} : \{kw_j\} \xrightarrow{describes} doc_i$$

This assignment can be done manually or automatically. **Manual indexing** means that people, skilled as natural language users and perhaps with expertise in the domain of discourse, have read each document (at least cursorily) and selected appropriate keywords for it. **Automatic indexing** refers to algorithmic procedures for accomplishing this same result. Because the index relation is the fundamental connection between the users' expressions of information need and the documents that can satisfy them, this simply stated goal – "Build the *Index* relation" – is at the core of the IR problem and FOA generally.

1.5.1 Automatically Selecting Keywords

We begin by considering the document at its most mechanical level, as a string of characters. Our first candidates for keywords will be **tokens**, things broken by **white space**. That is, each token in the document could be considered one of its keywords.

How good is this simple solution? Suppose users ask for documents about CARS and the document we are currently indexing has the string CAR. It seems reasonable to assume that users are interested in this document, despite the fact that the query happens to contain the **plural** form CARS while the document contains the singular CAR. For many queries we might like to consider occurrences of the words CAR and CARS, or even RETRIEVAL and RETRIEVE, as roughly interchangeable with one another; the suffixes do not affect meaning dramatically. And of course our problem doesn't end with plurals; we could make similar arguments concerning past-tense -ED endings and -ING participles.

This simple solution also depends too much on where spaces occur. Consider the German noun GESCHWINDIGKEITSBESCRANKUNG, corresponding to the English phrase SPEED LIMIT. In many ways, the fact that English happens to put a white space between the words while German does not is not semantically critical to the meaning of these descriptors or the documents in which they might occur. Such **morphological** features – used to mark relatively superficial, surface-structure features (such as tense or singular versus plural) – can be considered less important to the meaning. And differences between German and English are trivial when they are compared to Asian texts, where the relationship between *characters* and *words* is radically different.

What about **hyphenation?** Use of the word DATABASE, the phrase DATA BASE, and the hyphenated phrase DATA-BASE is highly variable, depending on author preference and current practice at the time and place of publication. Yet we would hope that all occurrences of any of these tokens would be treated as references to approximately the same semantic category. Similarly, we hope that the end-of-line hyphenation

(breaking long words at syllable boundaries) would not create two keywords when we would expect only one. But simply adding "-" to the set of white space characters defining tokens would make CLINTON-DOLE and A-Z keywords, too!

Hyphenation is concerned with the situation in which a potential keyword is broken up by punctuation; what about those situations where a space also breaks up a semantic unit? SPEED LIMIT seems semantically cohesive, but what algorithm could distinguish it from other **bigrams** (consecutive pairs of words) that happen to occur sequentially? The problem only becomes that much more complicated if we attempt to consider longer noun phrases like APPLES AND ORANGES or BACK PROPAGATION NEURAL NETWORK, let alone more complicated syntactic compounds such as verb phrases, clauses, or sentences. Identifying phrases is an important and active area of research from the perspectives of both IR and computational linguistics.

Summarizing, we will take a token to be our default keyword because this is straightforward. More sophisticated solutions will handle hyphenation, multiword phrases, subtoken stems, and so on (cf. Section 2.3.1).

1.5.2 Computer-Assisted Indexing

The field of library science has studied the manual process of constructing effective indices for a very long time. This standard becomes a useful comparison against which our best automatic techniques can be compared, but it also demonstrates how difficult comparison will be. There are data, for example, that suggest that the capacity of one person (e.g., the indexer) to *anticipate* the words used by another person (e.g., a second indexer or the query of a subsequent user) is severely limited [Furnas et al., 1987]; we are all quite idiosyncratic in this regard. The lack of interindexer consistency among humans must make us humble in our expectations for automated techniques.

But manual and automatic indexing need not be viewed as competing alternatives. In economic terms, if we had sufficient resources, we could hire enough highly trained catalogers to carefully read every document in a corpus and index each of them. If we couldn't afford this very expensive option, we would have to be satisfied with the best index our automatic system could construct. But if we have enough resources to hire one or two human indexers, what tools might we give them that would make the most effective use of their time?

We seek methods that *leverage* the editorial resource, in the sense that this manual effort does not grow as the corpus does. How might editors and librarians guide an automatic indexing process? What information should this computation provide that would allow intelligent human readers the assurance of a high-quality indexing function? Chapter 7 will discuss ways that editors can **train** machine learning systems, and a number of analyses that are of interest to editors will be mentioned, especially in Chapter 6.

1.6 FOA versus Database Retrieval

Within the field of computer science, the subfields of databases and IR are often closely aligned. Databases have well-developed theoretic underpinnings [Abiteboul et al., 1995] that have generated efficient algorithms [McFadden and Hoffer, 1994] and become the foundation for one of the most successful elements of the computer industry.

Both databases and search engines attempt to characterize a particular class of queries by which many users are expected to attempt to get information from computers. Historically, database systems and theory have been perceived as central to the discipline of computer science, probably more so than the IR techniques that are the core technologies for FOA. Things may be changing, however.

The general public's discovery of the Internet and subsequent interest in search engines like Alta Vista, InfoSeek, and Yahoo! suggest that many users find value in the lists of Web pages returned in response to searches. These search engines are clearly doing an important job for many people. It is also a quantitatively different job from organizing their address book (or record collection or baseball statistics) *databases*. How are IR and database technologies to be distinguished?

To make the distinctions more concrete, let's imagine a particular information need and think about how both a database and a search engine might attempt to satisfy it. An example query might be as follows.

QUERY 3 *What is the best SCSI disk drive to buy?*

TABLE 1.1 Hypothetical Database

Model number	*Manufacturer*	*Vendor*	*Size (GB)*	*Interface*	*Price ($)*	*Speed (msec)*
123	Seag	J&R	2.4	SCSI	162	12
123	Seag	Fry	2.4	SCSI	159	12
456	Metrop	A&B	2.5	IDE	∅	12.5
789	Seag	J&R	1.5	EIDE	121	10.5
...						

In the case of databases, strong assumptions must first be made about *structure* among attributes of individual records. Good database design demands that the fundamental elements of data, their format, and logical relations among them be carefully analyzed and anticipated in a **logical data model** long before any data are actually collected and maintained within a physical implementation. These assumptions allow specification of a syntax for the query language, strategies for optimizing the query's use of computational resources, and efficient storage of the data on physical devices.

Now let's assume that a logical data model has been constructed and that a large catalog of information from various hard drive manufacturers and vendors has been collated. We will also make the larger and problematic assumption that the users can translate the natural language of Query 3 into the somewhat baroque syntax of a query language such as SQL. The result of the database search might look something like Table 1.1.[†]

NLP for databases

Creating an example relation like this and populating it with a few instances is simple, but performing the necessary data modeling, collating the data from all of the manufacturers and vendors, and keeping it all up to date are much more daunting tasks. If the database catalog is out of date or missing data from important vendors, users might leave the database badly informed.

Now let's imagine using a search engine on the same query. When run against a UseNet news search engine like DejaNews, this query results in the retrieval shown in Figure 1.8 with the most highly ranked posting shown in Figure 1.9.

Users of this search engine will read about many issues *related* to hard disks, some of which may be *relevant* to their particular situation.

Matches 1-20 of 726 for search: **best SCSI disk**

	Date	Scr	Subject	Newsgroup	Author
1.	97/08/29	062	Re: IDE or SCSI?	comp.os.ms-windows.	hima@ur
2.	97/08/04	053	Re: switching from IDE t	comp.unix.unixware.	erthur@
3.	97/08/28	051	Re: Sun SCSI and Linux	comp.os.linux.hardw	tallpat
4.	97/08/03	050	Re: switching from IDE t	comp.unix.unixware.	Larry F
5.	97/08/02	050	Replace SCSI Disk on R3.	comp.unix.sco.misc	ericl@r
6.	97/08/21	049	FS: 486-100Mhz multimedi	misc.forsale.comput	tjloe@s
7.	97/08/11	049	SCSI problems	comp.os.linux.hardw	///ax 2
8.	97/08/03	049	Re: Which is Best: More	comp.os.linux.hardw	g r e @
9.	97/08/01	049	Best Scsi CDR for $300-$	alt.cd-rom	"Scott
10.	97/08/16	048	Re: IDE increases Interr	microsoft.public.wi	"Nolan
11.	97/08/16	048	FS:Micropolis 3243WAV dr	rec.video.marketpla	John Cr
12.	97/08/09	048	CDR Server --- Audio CD	hk.forsale	Mattgr
13.	97/09/01	047	comp.sys.apple2 Freq#8/1	comp.sys.apple2	nathan@
14.	97/08/31	047	Re: W95 Vs. MacOS Micros	comp.sys.mac.advoca	dfield@
15.	97/08/27	047	miro configuration utili	rec.video.desktop	Edy Ges
16.	97/08/05	047	[FOR SALE] Dual PowerPC	comp.sys.be.misc	"Frank
17.	97/08/04	047	Re: IBM OEM 0664-CSH and	comp.periphs.scsi	Chuck T
18.	97/08/01	047	error i kernel32.dll alt	no.pc	Jorunn
19.	97/08/12	046	SUN SPARC 20 FOR SALE.	comp.sys.sun.hardwa	power@i
20.	97/07/26	046	Re: Help: Replacing pri	comp.os.ms-windows.	Christo

FIGURE 1.8 Results of SCSI Search of UseNet

Subject: Re: IDE or SCSI?
From: helpful@urban.or.jp (me @ my humble abode)
Date: 1997/08/29
Message-Id: <3406f93a.50446926@nnrp.gol.com>
Newsgroups: comp.os.ms-windows.nt.setup.hardware,comp.windows.nt.misc
[More Headers]

On Thu, 28 Aug 1997 23:10:03 GMT, Michael Query <query@dpi.qld.gov.au> wrote:

>My question is, should I get another 2 Gb SCSI disk for putting the
>OS (NT 4.0 WS), software, etc on, or should I get an IDE disk for this?

Having played around with different conÆgsfor a while, I'd say go SCSI. I'd do that even if I had to get a second SCSI controller.

(You'll ``hear'' a lot of people arguing that IDE is good enough, but if you are after overall improved performance SCSI is best.)

my 2Y.

FIGURE 1.9 A Relevant Posting

For example, does the "best" qualifier in Query 3 mean lowest cost, maximum capacity, minimum access time, or something else? Can users choose between IDE and SCSI, or are they restricted to SCSI? Depending on what kind of users they are, some of the information retrieved may be

TABLE 1.2 IR versus Database Retrieval

	IR	*Database*
System provides	Pointer to data	Data item
User's query	General	Specific
Retrieval method	Probabilistic	Deterministic
Success criteria	Utility	(Correctness) Efficiency, User-friendliness, ...

immediately applicable to the purchase being considered, while other parts of it are better considered **collateral knowledge** (D. E. Rose, personal communication) that simply leaves users better informed.

A very different set of assumptions from those we made about the database system are necessary to imagine the search engine working. For example, who wrote these postings? Are they a credible source of good information; what is their **authority**? Well-trained database users should ask equally skeptical questions about the data retrieved, but rarely are authority, data integrity, and the like considered part of database analysis.

But the key assumption for our IR users is that they can "listen in" on this previous "conversation" and *interpret* the text that has been left behind as containing potential answers to the current question. The search engine is charged with retrieving textual passages that are likely to answer the users' questions. Once presented with these retrievals, FOA users have more humble expectations and are willing to do more interpretive work. Because FOA searches are often even less concrete than Query 3 and are issued by users simply trying to learn about a topic, *semantic* issues central to the interpretation of a textual passage and its context, validity, and so on are at the heart of the FOA enterprise.

Van Rijsbergen, p. 2, table 1.1 has summarized these issues along a number of dimensions by which IR and database systems can be distinguished, and several of these are duplicated in Table 1.2. Database systems are almost always assumed to provide data items directly. Search engines provide a level of indirection, a *pointer* to textual passages that contain many facts, hopefully including some of interest. The information need of the users is quite vague when compared to that of database

users. The search engine users are searching for information about a topic they don't completely understand. Typical database users have a fairly specific question, like Query 3, in mind. It might even be that the database is missing some data; for example, the special null value ∅ in Table 1.1 shows that the price of the third disk drive is not known. Even in this case, however, the database system "knows that it doesn't know" this information. FOA queries are rarely brought to such a sharp point; ambiguity is intrinsic to the users' expectations.

Because the queries are so general, an FOA retrieval must be described in probabilistic terms. If a particular hard disk's price is part of our database, we are certain, with probability = 1.0, of its value. Never would a database system reply with "This hard disk might cost about $300." As discussed in depth in Section 5.5, a search engine can use sophisticated methods for reasoning probabilistically, and available evidence might even allow it to be quite confident that retrieved items will be perceived as relevant. But never will we be entirely certain that a document is what users want; we can only have high confidence that it may be.

Finally, one of the problems in evaluating search engines is just what success criteria are to be used. We typically assume that information we get back from a database system is correct. (Try to find an ad for a database system that boasts, "Our system retrieves only right answers"!) One database system claims to be more efficient, cheaper, easier to integrate into existing code, and more user-friendly than others.

This list of ways that search engines might be distinguished from databases is far from exhaustive; Blair has proposed a more extensive analysis [Blair, 1984]. More recently, as search engine technology and WWW-inspired applications have both burgeoned, hybrids of databases and search engines have blurred the historical differences further. Some bases of database/search engine interaction are mentioned in Chapter 6.

Chapter 4 discusses the evaluation of search engines in great detail, but typically the bottom line is: Does the system help you? If you are writing a research paper, did this search engine help you find material that was useful in your research? If you are a lawyer preparing a case and you want to find every relevant judicial opinion, does the search engine offer an advantage over an equivalent amount of time combing

through books in a law library? Such squishy, qualitative judgments are notoriously difficult to measure, and especially to measure consistently across broad populations of users. The next section provides a quick preview of several precise measurements that have proven useful to the IR community but would not be found persuasive within the database community.

1.7 How Well Are We Doing?

Suppose you and I each build an FOA search tool; how might we decide which does the better job? How might a potential customer decide on their relative values? If we use a new search engine that seems to work much better, how can we determine which of its many features are critical to this success? If we are to make a science of FOA, or even if we only wish to build consistent, reliable tools, it is vital that we establish a methodology by which the performance of search engines can be rigorously evaluated.

Just as your evaluation of a human question-answerer (professor, reference librarian, etc.) might well depend on subjective factors (how well you "communicate") and factors that go beyond the performance of the search engine (does *any* available document contain a satisfying answer?), evaluation of search engines is notoriously difficult. The field of IR has made great progress, however, by adopting a methodology for search engine evaluation that has allowed objective assessment of a task that is closely related to FOA. Here we will sketch this simplified notion of the FOA task.

Omniscient relevance

The first step is to focus on a particular query. With respect to this query, we identify the set of documents *Rel* that are determined to be relevant to it.[†] Then a good search engine is one that can retrieve all and only the documents in *Rel*. Figure 1.10 shows both *Rel* and *Retr*, the set of documents actually retrieved in response to the query, in terms of a Venn diagram. Clearly, the number of documents that were designated both relevant and retrieved, $Retr \cap Rel$, will be a key measure of success.

But we must compare the size of the set $|Retr \cap Rel|$ to something, and several standards of comparison are possible. For example, if we are very concerned that the search engine retrieve every relevant document,

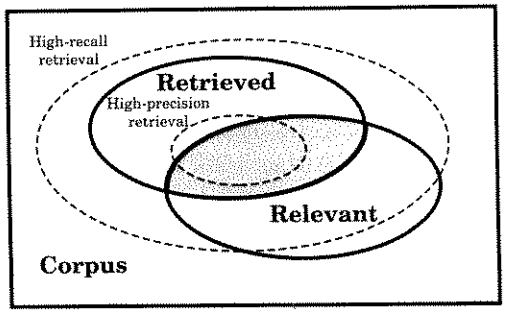


FIGURE 1.10 Comparison of Retrieved versus Relevant Documents

then it is appropriate to compare the intersection to the number of documents marked as relevant, $|Rel|$. This measure of search engine performance is known as **recall**:

$$\text{Recall} \equiv \frac{|Retr \cap Rel|}{|Rel|} \tag{1.1}$$

However, we might instead be worried about how much of what the users see is relevant, so an equally reasonable standard of comparison is what number of the documents retrieved, $|Retr|$, are in fact relevant. This measure is known as **precision**:

$$\text{Precision} \equiv \frac{|Retr \cap Rel|}{|Retr|} \tag{1.2}$$

Note that even in this simple measure of search engine performance, we have identified two legitimate criteria. In real applications, our users will often vary as to whether high precision or high recall is more important. For example, a lawyer looking for every prior ruling (i.e., judicial opinions, retrievable as separate documents) that is **on point** for his or her case will be more interested in **high-recall** behavior. The typical undergraduate, on the other hand, who is quickly searching the Web for a term paper due the next day, knows all too well that there may be many, many relevant documents somewhere out there. But the student cares much more that the first screen of **hits** be full of relevant leads.

Examples of high-recall and high-precision retrievals are also shown in Figure 1.10.

To be useful, this same analysis must be extended to consider the order in which documents are retrieved, and it must consider performance across a broad range of typical queries rather than just one. These and other issues of evaluation are taken up in Chapter 4.

1.8 Summary

This chapter has covered enormous ground and attempted to summarize topics that will be discussed in the rest of this text. Major points include:

- We constantly and naturally Find Out About (FOA) many, many things. Computer search engines need to support this activity, just as naturally.
- Language is central to our FOA activities. Our understanding of prior work in linguistics and the philosophy of language will inform our search engine development, and the increasing use of search engines will provide empirical evidence reflecting back to these same disciplines.
- IR is the field of computer science that traditionally deals with retrieving free-text documents in response to queries. This is done by indexing all the documents in a corpus with keyword descriptors. There are a number of techniques for automatically recommending keywords, but it also involves a great deal of art.
- Users' interests must be shaped into queries constructed from these same keywords. Retrieval is accomplished by matching the query against the documents' descriptions and returning a list of those that appear closest.
- A central component of the FOA process is the users' relevance feedback, assessing how closely the retrieved documents match what they had "in mind."
- Search engines accomplish a function related to database systems, but their natural language foundations create fundamental differences as well.

- In order to know how to shop for a good search engine, as well as to allow the science of FOA to move forward, it is important to develop an evaluation methodology by which we can fairly compare alternatives.

In this overview we've made some simplifying assumptions and raised more questions than we've answered, but that is the goal! By now, I hope you have been convinced that there are many facets to the problem of FOA, ranging from a good characterization of what users seek, to what the documents mean, to methods for inferring semantic clues about each document, to the problem of evaluating whether our search engines are performing as we intend. The rest of this book will consider each of these facets – and others – in greater detail. But like all truly great problems, issues surrounding FOA will remain long after this text is dust.

EXERCISE 1 How many computer science departments in the United States offer undergraduate classes in databases? In IR? How many graduate classes? How many journals or conference proceedings, associated with the ACM or IEEE, are published in each area?

TERMS INTRODUCED IN THIS CHAPTER

audience
author
authority
automatic indexing
bigrams
broader term
captions
closed vocabularies
collateral knowledge
context
controlled
conventional
corpus
database
describes
document
domain of discourse
electronic artifacts
exhaustive
finding out about
genre
high recall
hits
hypernym
hypertext
hyphenation
indexing
information need
information retrieval
IR
jargon
keywords
level of treatment
logical data model
manual indexing
marked alphabets
meta-data

morphological
narrower terms
natural language
natural
on point
open vocabularies
operators
passages
plural
precision
proxies
publication information
query
query language
recall
relevance feedback
relevant
retrieval method
retrieved
rhetoric
search engine
semiotics
simple queries
specific
spoken documents
success criteria
system provides
terms of art
tokens
train
transcripts
transitive
uncontrolled
user's query
users
vocabulary
vocabulary choice
vocabulary size
white space
word games
World Wide Web