5

Focusing the Light: Making Sense in the Information Explosion

The difficulty seems to be . . . that publication has been extended far beyond our present ability to make use of the record. The summation of human experience is being expanded at a prodigious rate, and the means we use for threading through the consequent maze to the momentarily important item is the same as was used in the days of square-rigged ships.

Vannevar Bush, "As We May Think"

Before the Internet, or even the widespread availability of digital computers, Vannevar Bush argued that society was creating information far faster than it could productively use it. Bush's 1945 observation—he was then director of the Federal Office of Scientific Research and Development, which directed the activities of over six thousand American scientists—led to the now-familiar metaphor, the "information explosion." Alvin Toffler popularized the phrase in his best-selling book, *Future Shock*, in 1970.

The word *explosion* sounds bad. People get hurt by explosions. Toffler used it to symbolize the difficulty he predicted we would have coping with rapid social and technological change and an overstimulating information environment. Information in the news, in our own areas of work, and even in entertainment is created far more quickly than we can consume it or make sense of it. For people drowning in information, Toffler believed, the explosion would be felt as a shock wave signaling arrival of the future.

According to historical accounts, of course, the information explosion is nothing new. We recognize that the volume of information available to people in the developed world has been increasing for hundreds of years. That perception comes from the ever-growing trail of written history, the growth of the literate population, and our easy access to information

provided by technology—printing presses, telephones, radio and television, computer networks, and electronic information-storage devices. Individuals perceive the growth of information differently—depending on their personal and vocational situations and the responsibility they feel for keeping current.

A simple way to limit the explosion would be to stop creating information. But slowing down the output of information is impractical and seems wrong-headed. Would we really want to cut back the scientific publication that accelerates the search for the discovery and cure of disease? Or curb publication of the daily news? In Western democracies, limiting publication conflicts with a fundamental freedom, freedom of the press. Nor, sensing a popular cause, do we recommend curtailing the creation of movies, television, and other forms of entertainment.

The problem with the information explosion is not really that there is too much information. We already realize that we cannot know or read everything; we need, each of us, only to keep up with the documents relevant to our particular interests. In modern society people specialize and consume individual information diets. We each want a certain portion of information about the world at large, a certain amount about national and local matters, and a good deal of specific information about our circle of friends, our interests, and our occupations. The real problem with the information explosion is that it presents us with two dilemmas: being overwhelmed by useless information and having difficulty finding quickly the specific information we need.

Organizing the Information Soup

When Vannevar Bush considered the problem of the information explosion, he proposed addressing it with what he called a "memex" device: "It consists of a desk. . . . On the top are slanting translucent screens, on which material can be projected for convenient reading. There is a keyboard, and sets of buttons and levers. . . . Books of all sorts, pictures, current periodicals, newspapers, are thus obtained" (1945:107).

This early sketch of an information desk was extended by J.C.R. Licklider in 1961 and later described as a networked computer work station.

> [The average person will have] his intellectual Ford or Cadillac—comparable to the investment he makes now in an automobile, or that he will rent one from a public utility that handles information processing as Consolidated Edison handles electric power. In business, government, and education the concept of "desk" may have changed from passive to active: a desk may be primarily a display-and-control station in a telecommunication-telecomputation system—and its most vital part may be the cable ("umbilical cord") that connects it, via a wall socket, into the procognitive utility net. (Licklider 1965:33).

Popular awareness of the information explosion has grown in tandem with the number of personal computers connected to the Net. The quantity of documents accessible by a personal computer—now reaching beyond the file system of one computer to systems all over the world—has increased by factors of millions.

One of Bush's influential ideas was to create direct links among the documents in the memex. Then, when someone reading an article comes upon a citation to another document, he or she could just press a button to jump directly to the article cited. Such linking is the defining characteristic of the hypertext systems developed in the 1980s and of the hypertext markup language that now permeates the World Wide Web. At its best, link-hopping is an efficient way to move between related articles and information sources.

Unanticipated by Bush was the explosion of publishing and self-publishing that now populates the Net. Because of this profusion, hopping across links through browsing or "surfing" is by now an impractical, unsystematic way to search for information. Although the average is skewed by the presence of index pages and big information sites, the average number of links leaving an individual web page is now about thirteen. Searching without a map can lead to interesting diversions but generally results in getting lost in cyberspace. Even completely automated web walkers, which hop across web pages at electronic speeds, now take days or even weeks to sweep through all the documents on the Internet.

Also implicit in Bush's approach was the assumption that the private organization of information by links would be augmented by enough librarians to help keep the world's knowledge organized. Here again, the proliferation of documents on the Net—which are constantly being changed, moved about casually, written, and deleted (and are usually unedited and unrefereed)—cannot be captured by existing library cataloging systems.

A useful technological method of finding relevant articles in the information soup of the Network is using the indexes and search services that retrieve documents according to the keywords they contain. Another response to the need to organize information on the web is the proliferation of pages listing links—sort of a "home brew" approach to particular topics. From a pragmatic perspective, these pages serve as fertile starting points for information foraging.

The Haystack Complexity Barrier

When we want to describe something difficult to find, we often use the metaphorical expression "as hard to find as a needle in a haystack." To get a sense of scale, I wanted to know just how difficult that task is. More specifically, I asked, how does finding a needle in a haystack—or, more generally, finding all of an unknown number of needles in a haystack—compare quantitatively with finding a set of relevant pages of information on the Internet?

There are many ways to search a haystack, and some of them provide analogies to our later discussion. One colleague suggested slyly that finding a needle is easy if you walk in the haystack with bare feet. Others suggested using magnets. If a haystack can be cut into parts, several searchers could use a divide-and-conquer strategy to search different substacks at the same time. In the simplest—and most arduous—approach, a person would pick up each piece of the haystack one at a time to see whether it is the needle or a blade of hay.

To answer the complexity question, I conducted a "field study" at the nearby Portola Feed Center. I assumed that hay in a haystack is packed at about the same density as it is in a bale. According to my simple observations, blades of hay in a bale are packed about ten to the linear inch and the average blade of broken hay in a bale is about eight inches long. Thus, a hundred blades of hay occupy a volume of about eight inches by a square inch and a cubic foot of hay contains about 21,600 blades. A cube-shaped haystack ten feet on a side would thus contain a little over twenty-one million blades of hay.

By comparison, near the end of 1997, the Internet contained about a hundred million web pages. If checking a blade of hay is comparable to

checking a page on the Net, the Net was then about as complex as five haystacks. What surprised me about this field study was not that the Net's size had exceeded the "haystack complexity barrier" but that it had apparently done so in mid–1997. At the growth rate of about a factor of ten per year, searching the Net (albeit in an automated fashion) will soon dwarf this proverbially difficult search problem.

Information Feast or Famine

> In the early 1990s, when browsers and search tools first appeared in the Internet, some of the researchers here told me I should consider finding another line of work because there would be no further need for librarians. But now they are calling for help in greater numbers, asking not only for documents but also for data, analysis, and ways to search the Web. People are overwhelmed by what comes back when they search the Net for information. It's feast or famine.
>
> Giuliana Lavendel, 1997

Most people use a search service to search the Net for information. These services perform much of the time-consuming work before the searcher ever contacts it. Search services use web walkers ("search engines") to sweep through the Net periodically, following links and keeping track of the documents or web pages they have visited before. The web walker records the words used in each document visited, then assembles them into an inverted index, which records the documents on which each word appears. The index is saved and later used to quickly match customers' queries for documents. Thus, the time-consuming work of sweeping the Net and constructing the inverted index is done ahead of time and does not delay the retrieval of results when a user requests a search.

When someone searching for a document designates the words that should appear in it, the search service uses the inverted index to find matching documents. For example, the disjunctive query *haystack complexity* returns a list of web documents in which either the word *haystack* or the word *complexity* appears. The conjunctive query *haystack + complexity* returns a list of web pages on which both words appear.

The breadth of a query is an indirect measure of how many matching documents or web pages one would expect to find. A broad query matches more documents than a narrow query. For example, the disjunctive query

haystack complexity search is broader than the query *haystack* because it matches more documents. Breadth may be quantified by counting the number of terms in a disjunctive query. For example, the disjunctive query *haystack search complexity* (i.e., documents containing the term *haystack* or *search* or *complexity*) has a breadth of three and would typically match more documents than the query *haystack search*, which has breadth of two.

Figure 5.1 shows how the number of documents retrieved from a large document collection varies with the breadth of a query. This curve reflects Giuliana Lavendel's "feast or famine" observation. For an example of the effect, suppose that we want to find a document that discusses the problem we are discussing here. Before writing this section, I connected to an online search service and started by asking for all documents that contain the phrase *haystack complexity barrier* verbatim. As this is a very narrow query, no matching documents were found. (This was not surprising, as I had invented the phrase while writing this chapter.) I then asked for all documents that contain that phrase or the term *search*. Over seventeen million matching documents were found, because the term *search* appears in so many places. To narrow down the results, I then limited the query, asking

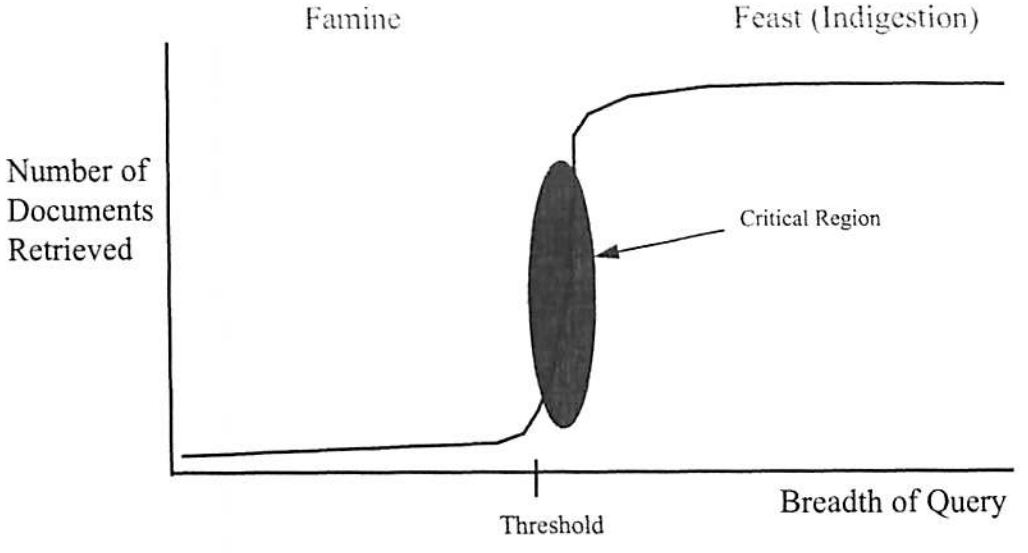

Figure 5.1
Phase Shift in Information Retrieval. For large document collections such as those found during an Internet search, a small change in the breadth of a query can result in a large change in the number of documents retrieved.

for all documents containing the three words *haystack, complexity,* and *search* in any order. The system found and ranked about fifty thousand documents containing at least one of these words. The top-ranked document was about managing complexity and included all the query terms, because it used the metaphor of finding a needle in a haystack. I then narrowed the query again by substituting the verbatim phrase *complexity barrier* for the word *complexity*. This change reduced the number of hits from fifty thousand to twenty-nine. The top-ranked document was an issue of an in-house magazine published by Digital Equipment Corporation about a line of computers. Thus, with only small changes in the query, the number of documents retrieved shifted from seventeen million to zero to fifty thousand to twenty-nine—fluctuating between feast and famine.

As suggested by this example, there are various ways to broaden or narrow a query by including certain terms or by requiring that they appear together, within a certain distance of each other, and so on. Short of running the query, we cannot determine the effects of particular changes with any precision.

This unpredictability and the extreme variation in the number of documents returned with small variations to the query indicate a phase shift in a search process; the two phases are the feast and famine in the number of documents returned. The term *phase shift* comes from physics and describes the sudden result of a small change, such as when a small decrease in temperature near the freezing point of water causes it to shift between liquid and solid phases. A point at which a transition occurs is called a *critical point*. A sharp transition from one phase to another at the critical point is called a *threshold effect*, and the region of rapid growth near a critical point is called the *critical region*. The shape of the graph in figure 5.1 is characteristic of a phase shift and is sometimes referred to as its *signature*.

Two standard measures of performance for information-retrieval systems are *precision* and *recall*. Precision is a measure of whether the documents returned by the process are relevant, and recall is a measure of whether all relevant documents are found. A deep problem of information retrieval is that near the critical region using a simple word-matching approach to retrieving documents forces extreme trade-offs between precision and recall. A broad query overwhelms the searcher with a flood of documents. However, though narrowing the query reduces the

number of documents returned and increases precision, it sacrifices recall, and relevant documents may be missed. The problem is actually worse than that. Even at the right-hand side of the phase transition where the searcher is deluged with documents, many relevant documents may be missed owing to a mismatch between the vocabulary of the query and the vocabulary used in the documents. Thus, the conventional tools for retrieving relevant and useful information are deeply flawed.

The feast-or-famine signature of a phase shift in information retrieval is more of a concern for some sensemakers than for others. The casual browser may be satisfied with the results of a single probe if most small subsets of documents retrieved include at least one relevant document. On the other hand, professional sensemakers searching for rare information face the challenging and time-consuming task of crafting queries to steer their search. Sometimes it is not clear until we begin that we are looking for a needle, that is, that the information will be difficult to find. Recognizing that users feel overwhelmed by the feast part of the feast-or-famine problem, some network search-tool providers hide the problem by not showing them the large number of hits actually resulting from their search.

This precision/recall dilemma is a classic example of the difficulties of coping with complexity. Knowledge-based systems are computer systems that solve complex problems by using representations of knowledge to guide their search for solutions. In the design of knowledge-based systems, knowledge is the key to coping with complexity. What we need to effectively mine the critical region of figure 5.1 is a way to identify the relevant documents more exactly. Such knowledge for identifying documents would enable a search service to pluck out the needed documents without deluging the searcher with materials that match a query for accidental reasons.

Technology for Making Sense

In the late 1990s, the number of documents on-line grew by a factor of ten per year. This growth was fueled by the rapid start-up of new commercial and academic servers, the increased ease of posting web documents on popular service providers like America Online, and by the international expansion of the Internet. Even so, much on-line information—such as that provided by the on-line news services or posted on the growing body

of private corporate intranets—never gets indexed. As digital rights technology is more widely integrated into the Net infrastructure, many more documents containing potentially high-quality information will be available on the Net for a fee.

Meanwhile, users' experiences with the feast-or-famine phase shifts of search services are creating a pushback from the Internet edge. Their frustration leaves many searchers feeling that the Net is an unreliable source. To the casual user of search services, the growth of information looks like an information explosion caused by network technology.

Although people may perceive technology as the cause of the information explosion, the most practical solution is also based on technology. It requires us to recast the problem from retrieving information to making sense of information.

Today's Leading-Edge Sensemakers

Within the increasing population of computer users there is a growing subset who spend a large fraction of their work life making sense of on-line information. These are people whose work requires them to sift through large quantities of data to understand something. Business analysts, who develop plans and strategic visions for new and established businesses, are sensemakers, as are analysts in government intelligence agencies and scientific leaders, especially those who work on multidisciplinary problems. Information specialists in libraries, policy analysts in think tanks and other information centers, and reporters and news analysts are all sensemakers. Trial lawyers looking for relevant cases are sensemakers. Patent attorneys looking for related patents and examining records of depositions are sensemakers. Students in college preparing reports on the material they are learning—as well as their professors conducting research in various fields of learning—are sensemakers. Even people who organize their e-mail into folders or devise bookmarks for the web are sensemakers. As they synthesize information from multiple sources, today's sensemakers have to extend their reach as more and more information becomes available on-line.

Sensemaking may seem like a solitary activity analogous to the lonely work of a scholar who spends years in the dimly lit archives of an old library. Yet sensemaking involves not only solitary individuals but also teams of

collaborating problem solvers. The work of electronic sensemaking ranges widely—from small temporary points requiring rapid assessment to recurring and complex problems that may take weeks or months to solve.

Private and governmental research and consulting organizations often include departments or pools of experts with different specialties. When a client organization requests information on a topic, a task force drawing on specialists from the different departments is formed. The specialists compile information from different sources and try to make sense of it. Figure 5.2 shows a model of collaborative sensemaking. The process begins when a client requests a report on a topic. The first stage defines the task, establishing its scope and addressing any ambiguities in the initial request. A task force is then selected from the available experts. Next, the task force meets to identify potential sources of information and to brainstorm the questions and sketch out the organization of the report. During this part of the process, members begin to find relevant documents and divide the work.

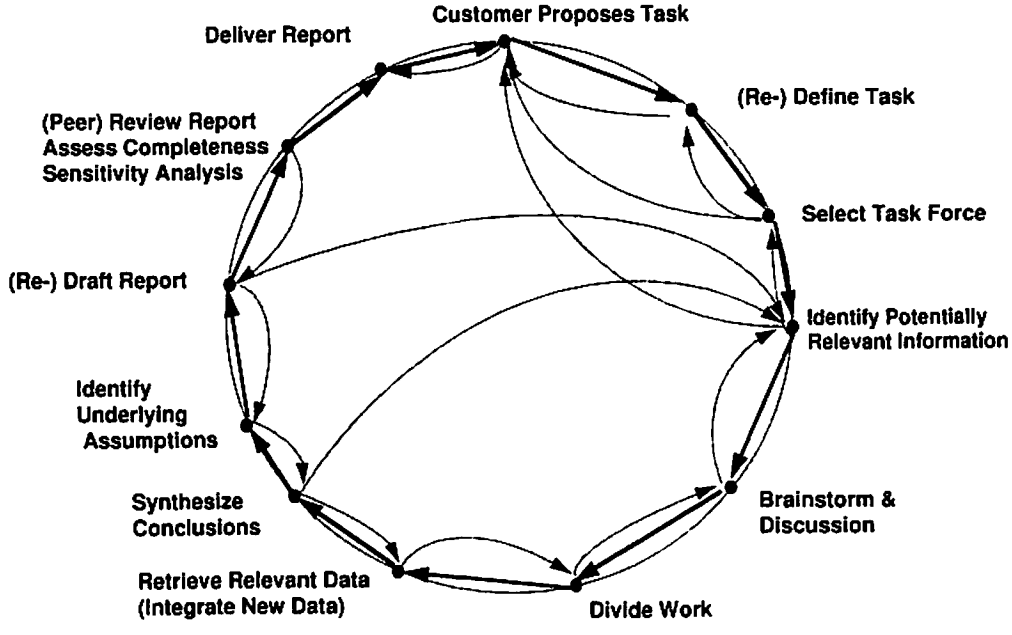

Figure 5.2
The Cyclic Work of Sensemaking. The graph illustrates the typical steps in a professional and collaborative sensemaking task, starting with a request by a customer who needs information and concluding with delivery of a report. The inner arrows in the cycle show that the task may involve feedback and looping. For simplicity, only a few of the many possible feedback loops are shown.

Once the work is divided, members of the task force begin to work in parallel on different parts of the task, pulling together bits of information and beginning to synthesize conclusions. Sensemaking about trends and generalizations often requires analysts to project the future and challenge each other to articulate sound assumptions on which to base projections. As the draft report takes shape, they may obtain peer reviews, which may challenge the report on the grounds of completeness or overdependence on unfounded assumptions. The cycle ends when the report is delivered to the client, although in many settings, such as the publishing of a newsletter, the cycle starts up again for the next edition.

The activity model of a collaborative team shown in figure 5.2 is not a perfect fit to all sensemaking situations. The work of any particular sensemaker or team of sensemakers depends on the social and institutional context. Some sensemakers work largely alone or in a loosely-connected network of colleagues. In a scientific setting sensemaking takes place when a group of colleagues and students turn their attention to writing a joint paper. A person writing a paper alone is also a sensemaker but collaborates with others only indirectly: by using others' published materials and in the peer review that takes place before publication. Some sensemaking tasks are so brief that they result in almost no written record. In other, more mature or seasoned sensemaking activities the explicit identification of underlying assumptions is a salient characteristic. The cycle described illuminates the range of sensemaking activities that are integrated into the reading, writing, and analysis of documents.

External Cognition

Psychologists who study how people work with knowledge or information pay close attention to their use of external representations: that is, to their writings and drawings on computer screens, paper tablets, or blackboards. They use the term *external cognition* for the human information processing that combines internal cognition with perception and manipulation of external symbols. We create, use, and share external representations in ways that augment our mental work.

I recently encountered a very effective example of a designed external representation when I was a member of a planning team preparing a proposal

for a new business. We presented our proposal for review to a corporate oversight group using spreadsheets that set forth the financial aspects of the plan in a specified format. During the meeting, I was struck by the efficiency with which members of the oversight group flipped quickly to particular pages and stepped through the columns of numbers. Their probing of our plan followed well-practiced lines: Why do you believe you can hire people so rapidly during this quarter? Do these expense figures account for the staggering of employment start dates during the period? Why do you expect the income to climb so rapidly during this period? For them the formatted spreadsheet was a familiar external representation that made it easy to find certain kinds of information. Their effective use of the document depended on the way it grouped together exactly the information they needed.

In crafting external representations of a task a sensemaker typically gathers information from many places. Making sense begins with the selection and organization of the information to be used for making a decision. External representations help the group or individual both to figure out the problem and communicate the solution. In the process initial scribbles and informal notes evolve into a formal report. This point in the work invites a useful twist in terminology—referring to the external representation itself as the *sense*—the product of the analytical process. Viewed in this way, sense is not just an internal understanding. In writing a report or crafting a representation, sensemakers are literally *making the sense.*

Query-free Retrieval

The industries growing up around the World Wide Web have brought us many generations of browsers and integrated suites of tools. Yet the process of developing powerful sensemaking tools that scale up to the rapidly increasing amount of information available is still in its infancy. In this section, we consider several leverage points for producing new, more powerful generations of sensemaking tools—places where technology can make a real difference in augmenting human sensemaking.

On-line sensemakers often start with many terabytes of on-line information. However, even using external representations, sensemakers can see or manipulate only a few pages of information at once. The ability to lever-

age the power of computers depends on designing external representations that have powerful affordances for sensemaking (like the spreadsheets designed to visually present the strategic analysis of a business proposition). How can we design sensemaking systems that, by manipulating only a few pages of writing, can give us the computational leverage to make sense of terabytes of information?

One innovative approach to the problem is designed to retrieve information or documents without creating a query. In *query-free retrieval* a system for working on-line is integrated with an information-retrieval system. In one application developed at Ricoh Silicon Valley, a diagnostic system named Fixit is used to fix printers and copiers. Besides the knowledge-based reasoning stored in its software, Fixit has automatic access to a database of maintenance manuals. To access a manual, a technician using the system need not type in a query or refine it to retrieve information about a specific problem. Fixit "knows" the context the user is working in and can offer relevant references to those portions of the manuals containing information on the diagnosed fault in the particular type or model of equipment. The essential key to automatic query generation is thus a system that has a detailed map of the topic of interest and a way of discovering what a technician needs by monitoring what he or she is doing in the diagnostic process.

The help systems for programs on personal computers make use of the same basic idea. Query-free retrieval has also been used in several projects at Apple, where a computer system retrieves on-line information for users based on the work they are doing. Some information-retrieval systems use a particular form of query-free retrieval called *relevance feedback*, in which the system retrieves additional documents from a repository whose word-usage profile most closely matches a test set. In chapter 2 we described how a PDR system might generate automatic queries from highlighted phrases or digital ink markings that active readers make on a digital document.

Sense Maps and Snippets

Imagine that a sensemaker is assembling some notes on the telecommunications industry in preparation for writing a report. He or she might use any number of possible ways to organize the data—writing separate ideas

and pieces of information in different regions of a page or structuring them with an outlining tool. Suppose that the sensemaker decides to make an outline to make sense of the various parts of the industry. The outline so far created is as follows:

I. U.S. Telecommunications Industry
  A. Industry Structure
    1. Regional Structure
    2. Forces for change
  B. Technological Trends.

Suppose further that while the sensemaker is editing the outline, the computer system is carrying out a search for relevant documents, using the outline to drive query-free retrieval. The system could open another window on the sensemaker's screen to list the retrieved documents, each of which relates in some way to the topics listed in the outline.

In the past few years text-processing systems capable of carrying out such tasks as writing document summaries have been created. They sometimes use the term *snippets* to refer to small chunks of text roughly the size of paragraphs. Snippets are bits of a document one might snip out with a pair of scissors. They are small segments on a single topic.

As the work of sensemakers is fundamentally compositional—sensemakers *make* sense—we have to ask: What are the units out of which they make it? Reports are too big for this purpose; sensemakers do not assemble an argument out of whole reports. The useful size for units of composition is the snippet. Consider, for example, the following portion of a snippet.

> The U.S. telecommunications industry is changing to create excess cellular and satellite capacity. This excess creates an opportunity for third parties . . .

We see looking back at the outline that this snippet seems most relevant to section I.A.2, which is about forces for change. How could the outline itself help us determine the most focused place for the snippet? Consider section I.A.2 again. Even without a more complete outline, we know from the structure of the outline that this section is not just about forces for change in general. It is about forces for change in the industry structure of the U.S. telecommunications industry. It is clearly not about regional structures or technological trends, which are covered in separate sections.

This picking apart of the outline into topics suggests a way to use its structure to target the mapping of snippets to the sense of the document. At any level of the outline the topic is essentially the conjunction (or logical additive-AND) of itself with the topics of its "parents" at all higher levels of the outline. As much as possible, entries at the same level should represent mutually exclusive (logical exclusive-OR) topics.

We use the term *sense map* to refer to an external representation that maps the snippets retrieved to the parts of an evolving sense document. As the sensemaker edits the sense—adding information, reordering the outline, or combining or splitting topics—the information search system is invoked, recomputes the set of relevant snippets, and presents them for possible incorporation.

In this way, the sense map is an artful way to combine the different compositional and retrieving aspects of making sense. In the community of researchers who have become interested in sensemaking, this combination is summarized by the following pseudoequation:

Sensemaking = Reading + Retrieving + Organizing + Authoring.

The equation says that the work of sensemaking is a process of finding and organizing relevant bits of information. Yet, although the electronic editors and browsers of the late 1990s are said to be integrated, they still split the work of sensemaking between separate tools for retrieving information and writing about it. And, working at the document level rather than the snippet level, they still require users to formulate queries.

Broadening Recall

The natural advantage of retrieving relatively long documents during a search conducted from a query accrues from the fact that most writers use a variety of equivalent phrasings to avoid producing a monotonous text. For example, an author may mention the *United States* in one passage, in another use the abbreviation *U.S.* or the adjective *American,* and in a third refer metaphorically to *Uncle Sam.* The retrieval of whole documents thus increases the chances that the words of the query will be matched to a relevant document.

This natural advantage does not apply so much to short document segments, in particular to snippets. Because snippet retrieval is less likely to

find multiple phrasings of a concept, searches need to use techniques that match documents according to meaning rather than just words.

Semantic matching can improve recall for whole documents too. Some techniques reducing the requirements for exact word matching are already routinely applied in information retrieval. For example, most retrieval systems use *word-stemming* systems to remove suffixes and prefixes to convert terms to a standard form for matching; example, the words *dreamer, dreaming*, and *dreams* would all be converted to the root word *dream.*

There are several other ways of broadening the basis of a match. One is to look for synonyms. Thus, a search for documents including the term *city* could be broadened to gather documents using words like *town, metropolis, suburb*, and so on. There are, however, some difficulties involved in routinely broadening retrieval by using synonyms. Frequently, whether two words are synonyms or not depends on their context. As a kind of trick question, I sometimes ask people whether the words *man* and *woman* are synonyms. The usual answer—"Of course not!"—reflects our understanding that gender differences often matter. However, in discussions in which the issue is a common humanity or legal rights, the terms *man, woman, human*, and *person* are generally synonyms. For example, if we are searching for court cases about a man being robbed in a car, it is probably also useful to find cases in which a woman is robbed in similar circumstances. *Synonomy*—and meaning more generally—are context specific.

There are a number of relationships that can be used to broaden information retrieval. Suppose, for example, that someone is writing an article about mammals eating fish. It would be useful to include in the retrieval an article about a cat eating a fish, even though the term *cat* is not a synonym of *mammal.* As a cat is a kind of mammal, there is a class relationship between mammals and cats. Or, similarly, suppose that someone is writing an article about governments and the taxing of citizens and that somewhere out in cyberspace is a snippet commenting on court rulings on citizens' taxes. In this case, even though *court* is not a synonym of *government*, nor is a court a special kind of government, the snippet may be relevant. A court is, after all, an arm of government. Thus, there are a variety of relationships between terms that can be used to loosen the requirements for exact matching in information retrieval.

Increasing Precision

I was interested in knowing whether anybody ever found any viruses that attack the malaria parasite. I have all of Medline titles and abstracts since 1966. I can search them for strings—all the usual—and there's even a matching vocabulary or thesaurus that will do some level of translation for equivalence. The trouble is that the nouns are there but the verbs are elusive. I can easily find articles in which viruses are mentioned and malaria is also mentioned, but none of them have to do with what I'm talking about. I have no way to capture "viruses attacking plasmodia." There are so many synonyms for that and I just get hundreds of articles that are about coincidental infections.

Joshua Lederberg, 1997

I once had a conversation with the geneticist, Joshua Lederberg, about his use of information retrieval. He was looking at new approaches for curing malaria. In the late 1990s, malaria killed 2.7 million people each year, mostly children. A new generation of vaccines was being tried, but with only partial success. Malaria has evolved in a way that keeps it one step ahead of the body's immune response system, shifting forms and sites of infection from the bloodstream, to the liver cells, to red blood cells. Lederberg believed it might be possible to use a "counterattack" approach based on viruses or other infectious agents that attack plasmodia, the parasites that cause malaria.

To this end he tried to construct a query to locate such agents with an electronic search service. He knew that retrieving articles about viruses was far too broad. He also found that retrieving papers that mention both viruses and malaria was inefficient, because there were numerous articles about people with malaria who also had secondary viral infections. He was overwhelmed by the large number of irrelevant documents that matched his query for what he considered accidental reasons. What he wanted was a way to tell the search system to retrieve all texts that mentioned the two words *malaria* and *virus* in a particular relationship. That idea is at the core of an approach called *schematic search*, which is intended to make retrieval more precise. Lederberg describes the concept as follows.

I did a little—I wouldn't even call it an experiment—a very hasty trial run. But I reckon there are only about thirty verb contexts that I would need to formulate and it would essentially solve my problem. Think of all the major relational connections between nouns. You know, inclusion and exclusion, modification, subtraction—it

doesn't take a great many of them. It didn't strike me as an impossible task to do this semantic conversion into a crude intermediate language. Perhaps that's a way that my problem might be solved. (1997)

Consider how this might work in the a less-technical example of a person studying the telecommunications industry. In this case, the sensemaker is looking for articles about non-U.S. companies buying a telecommunications company. Like Lederberg, the sensemaker wants to find documents or snippets in which a certain relationship holds among the words. Thus, the phrase "**Foreigners** *buy* telecommunications company" should match the following snippets:

MCI *to merge with* **British Telecom.**

NTT *considers buyout* of Motorola.

Siemens *increases holdings* in Deutsche Telekom.

In these examples, typographical differences indicate the parts of the snippets that correspond to the desired relationship among terms. Thus, the foreigners (boldface) in these snippets are British Telecom, NTT, and Siemens. The act of buying (italics) is expressed by the terms *merge, considers buyout,* and *increases holdings*. The telecommunications companies (underscore) are MCI, Motorola, and Deutsche Telekom.

The kind of semantic match in these examples is called a *schematic search*, where the initial phrase "foreigners buy telecommunications company" is used to create a schema that characterizes the required relationship. A schema indicates what kinds of terms or phrases can be used to fill in the relationship. Filling out a schema requires knowledge about the meanings of terms. Thus, to match the examples with non-U.S. companies requires knowing that British Telecom, NTT, and Siemens are the names of companies incorporated outside the United States. Matching the telecommunications company to the examples requires knowing that MCI, Motorola, and Deutsche Telekom are telecommunications companies. The relationship in this example is about buying a company. To match the examples requires knowing that merging, buying out, and increasing holdings are all ways of changing the ownership or control of a company.

It is possible to create computer systems that can work in the way these two examples suggest. Much of the research directed toward this development is taking place as part of the Message-Understanding Conference (MUC) sponsored by the Advanced Research Projects Agency (ARPA);

MUC evaluates empirical methods of extracting information from text. Such computer systems (or *knowledge-based systems*, as they are called) need to have encoded into them knowledge that enables them to perform the matches or, in the case of MUC, to extract information. This is precisely the kind of knowledge that would allow systems to work effectively in the critical region of the phase transition in the search process. Such knowledge is thus part of the critical leverage sensemaking systems need to cope with complexity and avoid the sharp trade-off between precision and recall—feast or famine—that plagues information retrieval.

The bad news, however, is that although encoding the knowledge for any given example is easy, the potentially enormous amount of knowledge of this sort needed to inform the search process for arbitrary topics makes this a massive task. The good news is that when semantic-matching knowledge isolates at least some of the pertinent articles on a given subject, the citations in the articles and in published indices can be used to locate related articles. Thus, given a starting point, related articles can be rounded up by following the citation links among the articles. The assumption that articles cite other articles works best for the scientific and scholarly literature. This is what happened to Joshua Lederberg when, months after he began his search, a colleague suggested searching for *Plasmodium* and *viruslike*. This netted a few articles, which then led to a treasure of other relevant articles.

Could the *viruslike* or *virus-like* term have been generated automatically by the retrieval-broadening process? There are many possible terms of similar meaning, including *viroid, quasi-virus, pseudo-virus, retrovirus, phage*, and *bacteriophage*. One search approach would treat generation and inclusion of these terms as domain-specific knowledge for broadening the retrieval; another would search the data base for terms that are variants of common roots. Both approaches have implications for system design. For example, proper use of such retrieval-broadening knowledge interacts with system elements for using synonyms, with the word-stemming system, and with other linguistic components of the matching software. Even if we solved the problem of how to generate a query using synonyms of *virus-like* as terms, we would find that including such terms in the search would intensify the precision problem by returning even more articles about secondary infections. In short, these approaches create a heightened need for semantic matching.

One approach to encoding the knowledge needed for schematic searching is to devise a system that allows individual sensemakers to encode such knowledge incrementally to increase the effectiveness of their work. The encoding for any particular sensemaker need only be complete enough to support the immediate task. However, because meaning is context-dependent, most sensemakers will need to maintain multiple minidictionaries of equivalences, perhaps even different dictionaries for different purposes. Furthermore, building a sensemaker's semantic-matching dictionary need not start from zero; he or she could begin with a small set of common relationships and then add to and tune it.

Attention Management

> What information consumes is rather obvious: it consumes the attention of its recipients. Hence a wealth of information creates a poverty of attention, and a need to allocate that attention efficiently among the overabundance of information sources that might consume it.
>
> Herbert Simon, as quoted by Hal Varian (1995)

Herbert Simon, the polymath economist and cognitive psychologist, has long studied how people use information in making decisions. Reflecting on the information explosion, he has often observed that human time and attention, not information, is the scarce resource. The biggest challenge is often how to allocate time to the most relevant information. From one perspective, this is just another corollary of his general notion of *bounded rationality*, the idea that people strategize to do make the best decisions that they can under the constraints of limited time and limited cognitive resources. As the quotation attests, there is a wealth of information but a poverty of attention.

Even in a sensemaking system that helps manage the information explosion by increasing both precision and recall, challenges to managing attention would remain. One impediment to incrementally building knowledge for semantic matching is that it requires sensemakers to perform two tasks simultaneously: making sense on the topic while tuning the search parameters—the semantic-matching knowledge of the system. The difficulty of doing so efficiently reflects the difficulties we have paying attention to several tasks at the same time.

Figure 5.3 illustrates the flow of snippets in a tool for sensemaking. As the user fills out an outline of the sense, snippets are retrieved from a repository and placed in the snippet arrival area. The shading of the snippets indicates how they map onto parts of the outlined sense. When a sensemaker selects a snippet, he or she can move it to the discard area (trash), to the snippet staging area, or to the sense.

An axiom of user interface design for collaborative tasks is that there should be a payoff to the user for any work he or she performs. When the user decides to discard a snippet into the trash, the system could ask for a specific a reason. Reasons for rejection might be, for example:

- *Bad source.* The sensemaker does not trust the information source. In this case, the system could ask the sensemaker to characterize the reason for rejecting the source in various ways, such as by the snippet's author, the news feed or publication that provided it, the person referenced in the snippet, or the snippet's genre.
- *Redundant data.* The sensemaker could have enough snippets of essentially the same kind, either because they are widely reported or because there are duplicates in the data base.

With the sensemaker's approval, the system could then intercept other snippets with the same characteristics and automatically toss them into the trash. Thus, as a payoff to the user for giving a reason to discard a snippet, the system provides greater assistance by discarding other snippets automatically.

In an ideal case, the snippet is directly useful and the user can incorporate it immediately into the sense document. In most cases, however, just copying the snippet into the sense document is a suboptimal approach. Usually it is better to encapsulate it into a digital object with other, similar meta-data. For example, the sensemaker could indicate that the snippet will be used as support for an argument, as a backup source or footnote to a section, or even to present a contrary opinion or counterexample. The value of such distinctions is that they enable the sensemaker to automatically analyze the completed sense report in terms of its use of data and sources.

The sensemaker may also decide to postpone action on a snippet. In this case, he or she could put the snippet into the snippet staging area with notations like the following:

- *Read later.* The sensemaker needs to read and think about the snippet but does not want to consider it right away; it has been selected as interesting but too complex for immediate use.

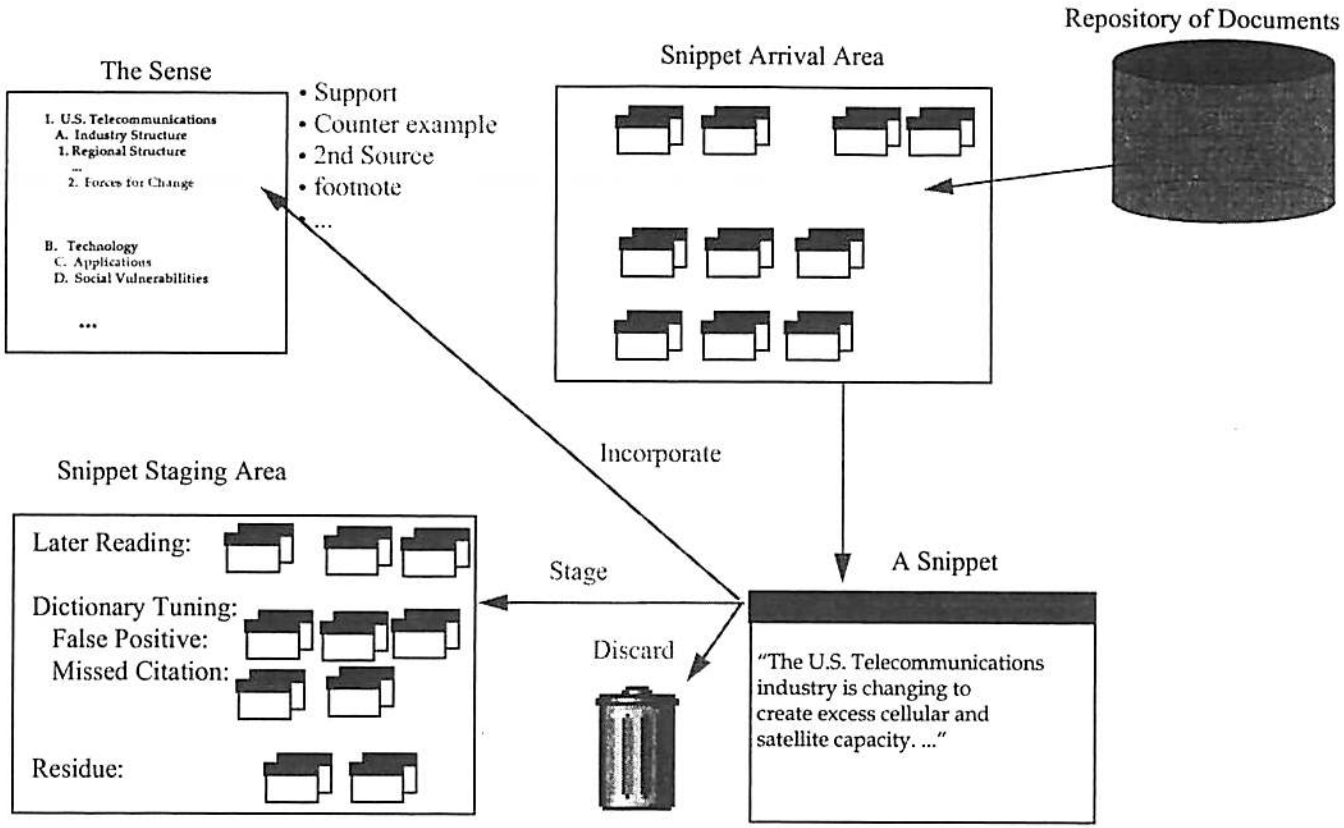

Figure 5.3
Managing Focus of Attention in a Sensemaking System. As the user fills out an outline of the sense (top left), the system searches a repository of documents for potentially matching snippets. The sensemaker can move a snippet to the discard area (trash), to the snippet staging area, or to the sense working area.

- *False positive.* This category means that the retrieval process has incorrectly collected the snippet. Saved as an example, the snippet can help guide or test the semantic matching by modifying the synonyms or other relations or fine-tuning the schematic search parameters.
- *Misfiled.* This category means that the snippet is interesting but probably belongs elsewhere in the sense document. Its appearance in the wrong location suggests that the sensemaker needs to tune the semantic-matching parameters of some other section of the report and use this snippet to guide the matching for it.
- *Residue.* The snippet challenges the basic categories of the developing sense and does not fit anywhere. The sensemaker needs to rethink the sense categories and then place this snippet where it belongs.

At the time of this writing, creating a sensemaking system like that described above is a-yet-unmet research challenge. Although the elements of such a system have been used in various information systems, they have not been tried all together in a system for sensemaking. Indeed, although the overall approach seems plausible, its effectiveness for sensemaking has not been demonstrated.

What can be said is that this proposal stands on fifty years of technology developed since Vannevar Bush first proposed the memex and addresses issues not then visible. It speaks in particular to how we might develop the knowledge needed to make the search for information more effective in the critical region. It also structures the overall task of sensemaking as an artful integration of reading, retrieving, organizing, and writing in a way that supports information retrieval from large document depositories without the need for formulated queries.

Reflections

The notion of bounded rationality causes one to reformulate what an information retrieval system should be in terms of benefit per unit time cost instead of precision and recall.

Stuart Card, 1997

The greatly increased amount of information now available on the Net—it has recently passed the haystack complexity barrier—has made the information explosion tangible for many people. Although thinkers like Vannevar Bush and J.C.R. Licklider anticipated the problem of the information

explosion several decades ago, their solutions for dealing with it were never tested—because the large on-line databases needed to do so did not exist in their time. Meanwhile, our experience with such systems has revealed deeper issues in using large collections of information that they never anticipated.

Now, although people expect ready information from the Net, what they usually experience is information feast or famine. Often they cannot even determine whether the information they need is on the Net. They face a threshold effect, either finding nothing or being deluged with matching but useless documents.

Suppose for a moment that we possessed sensemaking systems like those described in this chapter. Would they effectively solve the problem of the information explosion and the threshold effect? In attempting to answer that question, we are in a position not unlike that of Bush and Licklider, because we don't yet have the sensemaking systems to try out. We can, however, learn from a thought experiment.

At the core of the sensemaking proposal is the idea that query-free retrievals can be generated from the sense document a sensemaker creates. Although this approach offers the possibility of great cognitive leverage—manipulating two pages to make sense of terabytes—it also contains the seeds of a possibly dangerous flaw. The system as proposed essentially works by first determining and then amplifying the sense the sensemaker begins with.

A familiar phenomenon occurs when a group of writers passes around a draft of an article they are writing together. We call it the first-draft effect, because the first draft of the document has such a great influence over the final form of the document. If the first draft is fundamentally wrong in some way or blind to some issue, then later drafts are likely to be defective in the same way.

The same danger exists in sensemaking. If the first draft or first sense is wrong or lacking in some essential, the system and further writing will tend to amplify the error. As the sense is flushed out, it can become more and more difficult to think outside of the box. Of course, this problem is not limited to machine-assisted sensemaking. Reflective analysts have seen such bias effects in individual and collaborative sensemaking in which there is no machine amplification.

Perhaps the root of the problem lies in the standard measures of information retrieval—recall and precision. Especially with regard to amplifying bias, using these metrics strictly contributes to the effect. Maybe what is needed is a greater appreciation of the value of outliers and contrary information. Imagine, for example, a retrieval system that returns snippets in three categories: relevant (mainstream), secondary, and outliers (contradictory). Indeed, we might even be able to develop automatic means of representing relationships among the snippets and using such relationships to generate suggestions for modifying the sense.

Another intriguing possibility is the idea that sensemaking systems could provide the basis for a kind of accountability of sensemaking. Once during a visit to an intelligence organization, I heard about a conversation that took place as a senior analyst was reviewing a draft intelligence report written by a junior analyst.

Senior analyst: How did you conclude that we would approve building oil pipelines through _____? [a middle eastern country]
Junior analyst: My source was a speech the Senator gave at _____. [eastern college]
Senior analyst: Don't you know that what senators say in such public addresses is for public relations and not policy?

The example suggests that analysts learn a lot about evaluating sources and using them for reliable sensemaking. In a similar way, tools for sensemaking systems could record the use and disposal of information from different sources and the reasons why it is used or not used. One plausible benefit of such tools would be that they would record not only the conclusions of sensemaking but also crucial parts of the process of *making* the sense. Such an "audit trail" could become the basis of a descriptive practice of sensemaking for teaching. The records could be used by junior analysts learning by example or by senior analysts mentoring junior analysts about the rules of good sensemaking. Moreover, like the outside auditors called in to check a corporation's books and certify that it has used good accounting principles, outside sensemakers could use the record to check a sensemaker's product to certify that he or she followed good sensemaking practice.

An interesting tension that arises from this example is that crossing the line from implicit to explicit rules of interpretation can be fraught with

danger. Is the rule about public speeches valid for all public occasions? Does it apply only to senators, or is it about other public or private officials too? What exceptions are there to the rule? If the rule is not made explicit, then it cannot be acted on automatically, nor even passed on to colleagues easily. If a rule is only implicit, an audit of information potentially bearing on a decision would turn up sources ignored for no apparent reason. Clearly, formal sensemaking would challenge individuals and organizations to be explicit about their criteria and assumptions.

As we have seen before, the process of inventing the Net—including developing tools for finding and using the information on the Net—is also a process of shaping ourselves. We can design sensemaking systems that reinforce our biases, or we can devise ways to both leverage our access to information and challenge our interpretations of it. Sensemaking, like other uses of the Net, is a fundamentally social process. We have an opportunity to design not only technology but also appropriate ways of using it together.

The need for good accounting principles of sensemaking may become more crucial as more people rely more and more on the Net for information. A key, and potentially dangerous, characteristic of digital information is its intangibility and invisibility. As we increase the amount of information we obtain on-line, we risk becoming less familiar with and connected to the actual source of the information. Since we are not *there,* we are less able to use the clues and context of the situation to guide us in using the information. More than ever, we require expertise and care in combining information from multiple sources.

A good accounting practice for evaluating sensemaking could eventually become an important part of how society and individuals think about the effective use of the knowledge we generate and, especially, how to weigh our growing reliance on the Net to aggregate and distribute that knowledge.