that are used to perform tasks, so that designers of such instructions must be sensitive not only to their informational content but to their computational properties. In this light, they also suggest that one instructional representation of a device is very unlikely to be an optimal vehicle for supporting all user tasks: It may well be better to provide different representations of the same information, each tailored to particular tasks. In this sense, perhaps instructions should mirror and exploit the natural tendency, noted earlier, for users to form fragmentary mental models, with different fragments for different purposes.

In terms of theory, Bibby and Payne's findings lend support to the suggestion developed previously that mental models of a device that are formed from instructions may be computationally equivalent to the external representations of the device. This idea gives a rather different twist to the standard line that mental models are analog, homomorphic representations. It also supports a somewhat ironic, recursive suggestion concerning future research. The premise of this chapter is that understanding users' mental models can enrich our understanding of the use of cognitive artifacts. But perhaps an excellent way of further developing our understanding of the nature and genesis of mental models is to study the use of simple diagrammatic representations.

7 CHAPTER

Exploring and Finding Information

Peter Pirolli Palo Alto Research Center

7.1 INTRODUCTION

This chapter will discuss recent theories that address how people forage for information in their environment. The emergence of the global information ecology has created enormous pressures for users who seek useful information. Understanding how people adapt to these pressures has led to the development of information-foraging theory and the notion of users following information scent. Information-foraging theory is grounded in computational theories of human cognition and optimal foraging theories from biology. Applications of the theory to a novel browsing system (called Scatter/Gather) and the World Wide Web illustrate the utility of the approach.

7.2 MOTIVATION: MAN THE INFORMAVORE

The psychologist George Miller (1983) characterized humans as a species of *informavores:* organisms hungry for information about the world and themselves. We are curious creatures who gather and store information for its own sake, and we explore and use this wealth of information in order to better adapt to our everyday problems. We are not unique in this respect. There are other species that seem to be inherently driven by curiosity and a penchant for exploration and learning. We are, however, distinct in the extreme degree to which we have used technology to accelerate our evolution as informavores. This technology-boosted evolution is not without its tensions. Even though the amount of available information grows at an exponential rate,[1] our unassisted perceptual and cognitive capacities have not changed in any fundamental way since long before the invention of writing. Technological advances that increase the volume and flux of available information create pressures that must be met by technological advances that enhance our capacities for exploring, searching, and attending to that information. This chapter presents theoretical developments that aim to explain and predict how people will explore, search, and attend to the information environment, and how the information environment can be better shaped to people.

Much of the focus in this chapter will concern *information-foraging theory* and the concept of *information scent.* Information-foraging theory deals with understanding how user strategies and technologies for information seeking, gathering, and consumption are adapted to the flux of information in the environment. Information scent concerns the user's use of environmental cues in judging information sources and navigating through information spaces. These ideas evolved in reaction to the technological developments associated with the growth of globally distributed, easily accessible information, and the theoretical developments associated with the growing influence of evolutionary theory in the behavioral and social sciences.

7.2.1 Emergence of the Global Information Ecology

Information-foraging theory arose during the 1990s, coinciding with an explosion in the amount of information that became available to the average computer user, and with the development of new technologies for accessing and interacting with information. Hard-disk capacity grew 25% to 30% each year during the 1980s.[2] The number of Internet hosts grew at an exponential rate from 188 in December 1979 to 313,000 in October 1990 to 28,611,177 sites in March 2001.[3] Processing capacity continued to double every 18 months, inexorably following Moore's law.[4] Widespread deployment of improved scanners and fax machines meant that more documents were moving into, and flowing through, the digital world. High-resolution, high-performance, interactive graphics workstations enabled the development of novel information visualization and

1 See *www.sims.berkeley.edu/how-much-info/* for a variety of ways of measuring this growth.
2 James Porter of Disk/Trend, as cited in Toigo (2000).
3 See the Hobbes' Internet Timeline © at *www/isoc/guest/zakon/Internet/History/HIT.html*
4 Named after Gordon Moore, who observed that computer memory chip-capacity doubles every 18 months.

interaction techniques for large volumes of information. Whereas the average personal computer user of the early 1980s had access to perhaps dozens of files in local storage on their isolated machine, the average user in 1995 had access to about 275 million documents on the World Wide Web. It is estimated that the average user in 2001 had access to 525 billion Web documents.

The late 1980s witnessed several strands of human-computer interaction (HCI) research that were devoted to ameliorating problems of exploring and finding electronically stored information. It had become apparent that users could no longer remember the names of all their electronic files, and it was even more difficult for them to guess the names of files stored by others (Furnas, Landauer, Gomez, & Dumais, 1987). One can see proposals in the mid- to late-1980s HCI literature for methods to enhance users' ability to search and explore external memory. Jones (1986) proposed the Memory Extender (ME), which used a model of human associative memory (Anderson, 1983) to automatically retrieve files represented by sets of keywords that were similar to the sets of keywords representing the users' working context. Latent Semantic Analysis (LSA) (Dumais, Furnas, Landauer, Deerwester, & Harshman, 1988) was developed to mimic human ability to detect deeper semantic associations among words, like "dog" and "cat," to similarly enhance information retrieval.[5] This new work on ME and LSA, which was aimed at the "average" user, was contrasted with work in the "traditional" field of information retrieval in computer science, which had a relatively long history of developing automated systems for storing and retrieving text documents (vanRijsbergen, 1979). The 1988 CHI conference in which LSA was introduced also hosted a panel bemoaning the fact that automated information retrieval systems had not progressed to the stage where anyone but dedicated experts could operate them (Borgman, Belkin, Croft, Lesk, & Landauer, 1988). Such systems, however, were the direct ancestors of modern search engines found on the Web.

Hypermedia also became a hot topic during the late 1980s, with Apple's introduction of HyperCard in 1987, the first ACM Conference on Hypertext in 1987, and a paper session at the CHI 1988 conference. The very idea of hypertext can be traced back to Vannevar Bush's *Atlantic Monthly* article "As We May Think," published in 1945 (Bush, 1945).[6] Worried about scholars becoming overwhelmed by the amount of information being published, Bush proposed a mechanized private file system, called the Memex, that would augment the

5 Latent Semantic Analysis proved to be surprisingly successful in addressing psychological facts about human vocabulary learning (Landauer & Dumais, 1997)—an example of application-oriented research making a contribution to fundamental science.
6 See Buckland (1992) for a discussion of the developments that preceded Bush's publication.

memory of the individual user. It was explicitly intended to mimic human associative memory. Bush's article influenced the development of Douglas Englebart's NLS (oNLine System), which was introduced to the world in a tour-de-force demonstration by Englebart at the 1968 Fall Joint Computer Conference. The demonstration of NLS—a system explicitly designed to "augment human intellect" (Englebart, 1962)—also introduced the world to the power of networking, the mouse, and point-and-click interaction. Hypertext and hypermedia research arose during the late 1980s because personal computing power, networking, and user interfaces had evolved to the point where the visions of Bush and Englebart could finally be realized for the average computer user.

The confluence of increased computing power, storage, and networking and information access and hypermedia research in the late 1980s set the stage for the widespread deployment of hypermedia in the form of the World Wide Web. In 1989, Tim Berners-Lee proposed a solution (Berners-Lee, 1989) to the problems that were being faced by the Conseil Européen pour la Recherche Nucléaire (CERN) community in dealing with distributed collections of documents, which were stored on many types of platforms, in many types of formats. This proposal lead directly to the development of HyperText Markup Language (HTML), HyperText Transfer Protocol (HTTP), and, in 1990, the release of the Web. Berners-Lee's vision was not only to provide users with more effective access to information, but also to initiate an evolving web of information that reflected and enhanced the community and its activities.

The emergence of the Web in the 1990s provided new challenges and opportunities for HCI. The increased wealth of accessible content, and the use of the Web as a place to do business, exacerbated the need to improve the user's experience on the Web. It has been estimated that

- 65% of virtual shopping trips on the Web end up in failure (Souza, 2000),
- for every 1 million visitors, 40% do not return because of problems in Web site design, at a cost of $2.8 million (Manning, McCarthy, & Souza, 1998),
- Web site redesigns that failed to solve such problems were estimated to cost $1.5 million to $2.1 million in 1999 (Manning et al., 1998).

Such figures translate into an enormous loss of potential revenue and profit for a global industry expected to top $1 trillion in 2003.[7] This has created a strong

[7] Source: SearchEngineWatch.

demand for theory, technologies, and practices that improve the usability of the Web.

The usability literature that has evolved surrounding the Web-user experience is incredibly rich with design principles and maxims (Nielsen, 2000; Spool, Scanlon, Schroeder, Snyder, & DeAngelo, 1999), the most important of which is to test designs with users. Much of this literature is based on a mix of empirical findings and expert ("guru") opinion. A good deal of it is conflicting.[8] The development of theory in this area can greatly accelerate progress and meet the demands of changes in the way we interact with the Web. Greater theoretical understanding and the ability to predict the effects of alternative designs could bring greater coherence to the usability literature, and it could provide more rapid evolution of better designs. In practical terms, a designer armed with such theory could explore and explain the effects of different design decisions on Web designs before the heavy investment of resources for implementation and testing. This exploration of design space is also more efficient because the choices among different design alternatives are better informed: Rather than randomly generating and testing design alternatives, the designer is in a position to know which avenues are better to explore and which are better to ignore. Unfortunately, cognitive-engineering models that had been developed to deal with the analysis of expert performance on well-defined tasks involving application programs (Pirolli, 1999) had no applicability to understanding foraging through content-rich hypermedia, and consequently new theories were needed.

7.3 SCIENTIFIC FOUNDATIONS

This chapter presents two related strands of theory. These strands roughly correspond to the two aspects of Berners-Lee's vision for the Web: one strand dealing with the individual users' interactions with vast amounts of information, the other strand dealing with the phenomena that emerge from communities of interacting agents and knowledge. The connection among these strands is that they take an *adaptationist approach*. Users are viewed as complex adaptive agents who shape their strategies and actions to be more efficient and functional with respect to their information ecology. System designs (including hypermedia content) are similarly analyzed with regard to how they engage more functional

[8] This state of affairs in recognized in the maxim of one "guru" who claims that the only hard-and-fast principle is "it depends."

and efficient user behavior, and how they are adaptive with respect to user needs. By assuming many interacting complex adaptive agents, one may develop models that predict emergent phenomena that arise when many people produce and consume content, and when they interact with one another on the World Wide Web.

7.3.1 Influence of Evolutionary Theory: Adaptationist Approaches

The rise of adaptationist approaches in the study of human behavior during the 1980s deeply affected evolutionary biology as well as the behavioral sciences.[9] The approach gained currency in cognitive science during the 1980s as a reaction to *ad hoc* models of how people performed complex cognitive or perceptual tasks. At that time, models of cognition and perception were generally *mechanistic,* detailing perceptual and cognitive structures and the processes that transformed them.[10] The Model Human Processor (MHP) and Goals, Operators, Methods, and Selection rules (GOMS) (Card, Moran, & Newell, 1983) are cognitive-engineering examples that derive from this approach. The MHP specifies a basic set of information storage and processing machinery, much like a specification of the basic computer architecture for a personal computer. GOMS specifies basic task-performance processes, much like a mechanical program that "runs" on the MHP.

Around the same time that GOMS and MHP were introduced into HCI, there emerged a concern among cognitive scientists that mechanistic information-processing models, by themselves, were not enough to understand the human mind (Anderson, 1990; Marr, 1982). A major worry was that mechanistic models of cognition had been developed in an *ad hoc* way and provided an incomplete explanation of human behavior.[11] It had become common practice to cobble together a program that simulated human performance on some task and then claim that the program was in fact a theory of the task (Marr, 1982, p. 28). Anderson (1990, p. 7) lamented that cognitive modelers "pull out of an infinite grab bag of mechanisms bizarre creations whose only justification is that

[9] Adaptationism has been a controversial idea in the behavioral and social sciences because it raises questions regarding the fundamental nature of mankind and a host of moral issues. A fascinating account of this "opera" can be found in Segerstrale (2000).

[10] Mechanistic models of complex behavior, like problem solving, were an enormous advance in and of themselves.

[11] Note that Marr and Anderson are only more recent proponents of the stance.

they predict the phenomena in a class of experiments. . . . We almost never ask the question of *why* these mechanisms compute the way they do" (emphasis added). Miller, in his article about "informavores," commented on the incompleteness of the mechanistic approach by using the following analogy: "Insofar as a limb is a lever, the theory of levers describes its behavior—but a theory of levers does not answer every question that might be asked about the structure and function of the limbs of animals. Insofar as the mind is used to process information, the theory of information processing describes its behavior—but a theory of information processing does not answer every question that might be asked about the structure and function of the minds of human beings" (Miller, 1983, p. 112).

Roughly speaking, the adaptationist approach involves reverse engineering. Human behavior, and the human mind behind that behavior, have adapted to the environment through evolution. This approach leads one to ask *what* environmental problems are being solved and *why* cognitive and perceptual systems are well adapted to solving those problems. For instance, Anderson and Milson (1989) analyzed the particular mathematical functions that characterize forgetting from human memory. Memories tend to decay very rapidly if they are not used frequently or have not been used recently. The specific mathematical form of this decay function was shown to be optimal given the likelihood of information recurring in the environment and the assumption that the cost of searching memory increases with the number of items stored. Later, Anderson and Lebievre (2000) proposed that theoretical memory mechanisms would give rise to those functions. In other words, Anderson and Milson started by asking the question "what problem does memory solve" and assumed that the answer was "to recall events relevant to ones that reoccur in the environment." They then proceeded to analyze the structure of this environmental problem and proposed an optimal mathematical solution (a kind of engineering-design specification). Anderson later proposed a specific mechanistic design (a computer simulation of human memory) that satisfied this design specification.

The mind is a fantastically complex, cobbled-together machine, which has been incrementally designed by evolution to be well tailored to the demands of surviving and reproducing in the environment. The adaptationist approach recognizes that one can better understand a machine by understanding its function.[12] Children can figure out quite a lot about the operation of a VCR by

[12] This notion of a scientific observer trying to understand something on the basis of its design is called an *intentional stance* (Dennett, 1988). This is very close to Newell's notion of understanding a system at the *knowledge level* (Newell, 1982, 1993). When the design is assumed to be driven by evolution, it is called an *ecological stance* (Bechtel, 1985).

understanding its purpose with virtually no knowledge of its mechanism. Some cognitive scientists realized that human information processing was an extension of human biology and that "nothing in biology makes sense except in the light of evolution" (Dobzhansky, 1973).

In cultural anthropology, adaptationist approaches arose as a reaction to the lack of systematic theory in that area (Harpending, 1993), which is also common in many areas of HCI:

> Mainstream sociocultural anthropology has arrived at a situation resembling some nightmarish short story Borges might have written, where scientists are condemned by their unexamined assumptions to study the nature of mirrors only by cataloging and investigating everything that mirrors can reflect. It is an endless process that never makes progress . . . and whose enduring product is voluminous descriptions of particular phenomena. (Tooby & Cosmides, 1992, p. 42)

Adaptationist approaches grew in reaction to sociocultural researchers whose aim was no more than recording and taxonomizing observed behavior in an atheoretical manner.

7.3.2 Information-Foraging Theory

One strand of theory that addresses information-seeking behavior by individuals is information-foraging theory. Some essential ideas behind this theory include:

- *Exaptation of food-foraging mechanisms and strategies for information foraging.*[13] Natural selection favored organisms—including our human ancestors—that had better mechanisms for extracting energy from the environment and translating that energy into reproductive success. Organisms with better food-foraging strategies (for their particular environment) were favored by natural selection. Our ancestors evolved perceptual and cognitive mechanisms and strategies that were well adapted to the task of exploring the environment and finding and gathering food. Information-foraging theory assumes that modern-day information foragers use perceptual and cognitive mechanisms that carry over from the evolution of food-foraging adaptations.
- *Information scent.* In exploring and searching for information, users must use proximal cues—cues they can perceive in their local environment—to judge distal information sources and to navigate towards them. Underlined link text on Web pages or bibliographic citations in an online library catalog are examples of proximal cues that users use in this manner.
- *The economics of attention and the cost structure of information.* The late Nobel Laureate Herbert A. Simon observed that a wealth of information creates a poverty of attention and an increased need to efficiently allocate that attention. In an information-rich world, the real design problem to be solved is not so much how to collect more information, but, rather, to increase the amount of relevant information encountered by a user as a function of the amount of time that the user invests in interacting with the system. If a user can attend to more information per unit time, then the user's information processing capacity is increased, thereby amplifying cognition. The structure of the physical and virtual worlds determines the time costs, resource costs, and opportunity costs associated with exploring and finding information. People have limited attention and must deploy that attention in a way that is adaptive to the task of finding valuable information. Natural and artificial information systems evolve toward states that maximize the delivery of more valuable information per unit cost (Resnikoff, 1989). People prefer information-seeking strategies that yield more useful information per unit cost, and they tend to arrange their environments (physical or virtual) to optimize this rate of gain. People prefer, and consequently select, technology designs that improve returns on information foraging.
- *Relevance of optimal-foraging theory and models.* If information foraging is like food foraging, then models of *optimal foraging* developed in the study of animal behavior (Stephens & Krebs, 1986) and anthropology (Winterhalder & Smith, 1992) should be relevant. A typical optimal foraging model characterizes an agent's interaction with the environment as an optimal solution to the tradeoff of costs of finding, choosing, and handling food against the energetic benefit gained from that food. These models would look familiar to an engineer because they are basically an attempt to understand the design of an agent's behavior by assuming that it is well engineered (adapted) for the problems posed by the environment. Information-foraging models include optimality analyses of different information-seeking strategies and technologies as a way of understanding the design rationale for user strategies and interaction technologies.

[13] An *exaptation* is a feature that evolved as an adaptation to one kind of problem but then became an adaptive solution to another problem.

It should be emphasized that the development of information-foraging theory has proceeded by establishing a narrow but solid base, and that base is just starting to extend to a broader scope of phenomena. The initial focus has been on understanding how people work on relatively well-defined information-seeking problems, such as "find all documents that discuss GOMS." Within this scope, detailed models have been developed that make many strong or simplifying assumptions. Consequently, the models are surely wrong on some details, and they surely will need to incorporate additional complexities in order to deal with a broader range of phenomena. For instance, later, in the discussion of the detailed example in Section 7.4, models are presented that make the simplifying assumption that users maximize the number of relevant documents of a task. This happened to be true of the tasks that were studied to test the model (i.e., the instructions asked study participants to maximize the number of relevant documents found); in real-world tasks, however, one imagines that the initial set of documents retrieved will be useful, but redundant documents encountered later would have no value. The initial model can be modified to deal with this complexity and others by developing appropriate analyses of the utility of information for the user's task. The scientific bet is that the simple, strong models will provide comprehensible basic insights that can provide a foundation for tackling more complex problems and productively generating further insights. Like all analogies, the one between information foraging and food foraging breaks down at some point. For instance, information can be easily copied and distributed, but food cannot. Search in food foraging is constrained by locomotion in three-dimensional Euclidean space, but the information forager can move through information spaces in much more complex ways. Information-foraging theory draws many ideas from food-foraging theory, but many of those ideas must be modified in the process.

7.3.3 Optimal-Foraging Theory

Optimal-foraging theory (Stephens & Krebs, 1986) seeks to explain adaptations of organism structure and behavior to the environmental problems and constraints of foraging for food. Consider a hypothetical predator, such as a bird of prey. Its fitness will depend on its reproductive success, which in turn will depend on how well it does in finding food that provides energy. The environment surrounding this bird will have a patchy structure, with different types of habitat (such as meadows, woodlots, ponds, etc.) containing different amounts and kinds of prey. For the bird of prey, different types of habitat and prey will yield different amounts of net energy if included in the diet. Furthermore, the different prey types will have different distributions over the environment. For the bird of prey, this means that the different habitats or prey will have different access or navigation costs. Different species of birds of prey might be compared on their ability to extract energy from the environment. Birds are better adapted if

they have evolved strategies that better solve the problem of maximizing the amount of energy returned per amount of effort. Conceptually, the optimal forager is one that has the best solution to the problem of maximizing the rate of net energy returned per effort expended, given the constraints of the environment in which it lives.

Now consider an office worker or academic researcher facing the recurrent problems of finding valuable task-relevant information. The environment surrounding these foragers will also be patchy in structure and will contain different types of external media, such as books, manuscripts, or access to online documents with different kinds of interfaces and content designs. The different information sources will have different profitabilities, in terms of the amount of valuable information returned per unit cost in processing the source. Access (or navigation costs) to get the information will vary. The optimal information forager is one that best solves the problem of maximizing the rate of valuable information gained per unit cost, given the constraints of the task environment. In the information sciences, Sandstrom (1994) has suggested that optimal-foraging theory may successfully address the information-foraging behavior of academic researchers in a field.

Optimization models are a powerful tool for studying the design features of organisms and artifacts. Optimization models in general include the following three major components:

- *Decision assumptions,* which specify the decision problem to be analyzed. Examples of such information-foraging decisions include how much time to spend processing a collection of information or whether or not to pursue a particular type of information content.
- *Currency assumptions,* which identify how choices are to be evaluated. Information-foraging theory will assume information value as currency. Choice principles include maximization, minimization, and stability of that currency.
- *Constraint assumptions,* which limit and define the relationships among decision and currency variables. These will include constraints that arise out of the task structure, interface technology, and the abilities and knowledge of a user population. Examples of constraints includes the rate at which a person can navigate through an information-access interface, or the value of results returned by bibliographic-search technology.

In general, all activities can be analyzed according to the value of the resource currency returned and costs incurred, which are of two types: (1) *resource costs* and (2) *opportunity costs* (Hames, 1992). Resource costs are the expenditures of time, calories, money, and so forth that are incurred by the chosen activity. Opportunity costs are the benefits that could be gained by engaging in other activities, but are forfeited by engaging in the chosen activity. The *value* of information (Repo, 1986) and the *relevance* of specific sources (Saracevic, 1975; Schamber, Eisenberg, & Nilan, 1990) are not intrinsic properties of content (e.g., documents, Web pages) but can be assessed only in relation to the embedding task environment.

The use of optimization models should not be taken as a hypothesis that human behavior is actually optimal—as if the individual had perfect information and infinite resource to make decisions. A more successful hypothesis about humans is that they exhibit *bounded rationality* or make choices based on *satisficing* (Simon, 1955). Satisficing can often be characterized as localized optimization (e.g., hill climbing) with resource bounds and imperfect information as included constraints.

7.4 DETAILED DESCRIPTION: SCATTER/GATHER

Scatter/Gather (Cutting, Karger, & Pedersen, 1993; Cutting, Karger, Pedersen, & Tukey, 1992) is an interaction technique for browsing large collections of documents that is based on automatically identifying clusters of related documents. Studies of the Scatter/Gather system (Pirolli, 1997; Pirolli & Card, 1995, 1999; Pirolli, Schank, Hearst, & Diehl, 1996) illustrate the adaptationist approach embodied in information-foraging theory. A detailed analysis of the cost structure of the information environment was performed and models from optimal foraging theory were borrowed to develop predictions about the optimal strategies for users to undertake in interacting with Scatter/Gather. This analysis was then used to develop a detailed mechanistic computer program to simulate individual users. This computer simulation was, in turn, used to evaluate potential design improvements (Pirolli, 1998).

It is not unusual for corporate and e-commerce Web sites to contain tens of thousands of documents. Portals such as Yahoo! are far larger. Web designers for large sites face the difficult task of developing site organizations and links that, on the one hand, get people to the information they seek and, on the other hand, provide some overview of all the content on the site. One technique is to develop a hierarchically organized categorical structure, much like that used in library catalog systems (this is the main approach used in Yahoo!). Users interacting with such structures follow links that seem to lead to progressively more refined topics that match their information goals (Spool et al., 1999). In general, a hierarchically organized search is extremely efficient, and a general model of

search costs in such systems is presented in Section 7.5.1.[14] Web site designers face the daunting task of creating category structures that make sense to as many users as possible. One problem, however, is that even the same subject domains may be conceptualized in different ways by different types of people. In addition, the effort required to maintain a hierarchical directory structure by human means has a number of scaling problems. Increased person power is required as more documents are added to a Web site, as the rate of document change increases, or as the user base changes or becomes more diversified. Scatter/Gather (Cutting et al., 1992) addresses these issues by providing a reasonably fast automatic document-clustering technology that organizes the contents of a collection in a hierarchical manner that is continually modified in reaction to user interests.

Figure 7.1 presents a typical view of the Scatter/Gather interface. The document clusters are separate areas on the screen. Internally, the system works by precomputing a *cluster hierarchy*, recombining precomputed components as necessary. This technique allows the interactive reclustering of large document collections in reasonable amounts of time. The clustering in Scatter/Gather depends on a measure of interdocument similarity computed from vectors that reflect the frequency of words in each document (vanRijsbergen, 1979). The Scatter/Gather clustering method summarizes document clusters by *meta-documents* containing profiles of topical words and the most typical titles. These topical words and typical titles are also used to present users a summary of the documents in a cluster. Topical words are those that occur most frequently in a cluster, and typical titles are those with the highest similarity to a centroid of the cluster. Together, the topical words and typical titles form a *cluster digest*. Examples of these cluster digests are presented in each subwindow in Figure 7.1.

7.4.1 Task Analysis

The user may *gather* those clusters of interest by pointing and selecting buttons above each cluster in Figure 7.1. On command, the system will select the subset of documents in these clusters and then automatically *scatter* that subcollection into another set of clusters. With each successive iteration of scattering and gathering clusters, the clusters become smaller, eventually bottoming out at the level of individual documents. At any point, the user may select a set of clusters and request that the system display all the document titles in those clusters. These appear in a scrollable "titles" window. The user may select a title from this window

[14] An analysis of the physical card catalog system (Resnikoff, 1989) that used to be found in most libraries suggests that most were near-optimal in their arrangement.

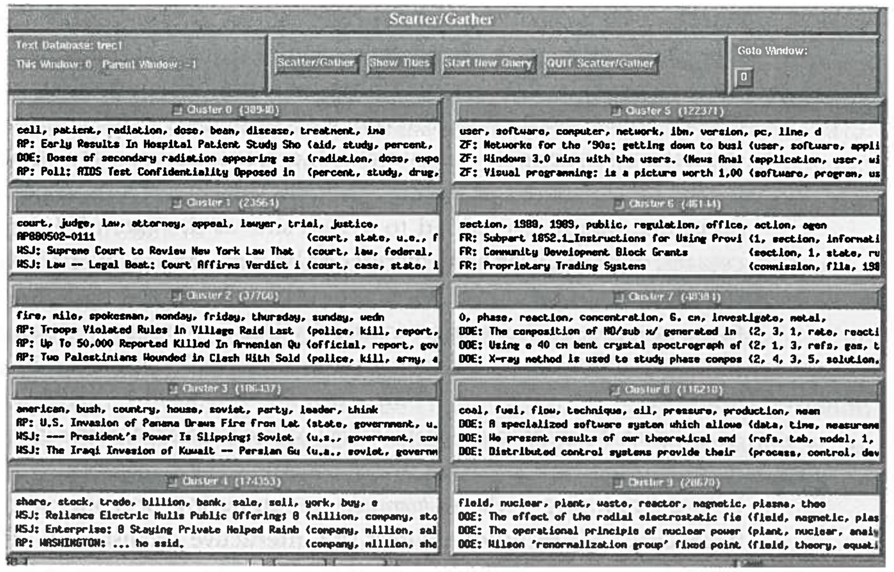

FIGURE 7.1 The Scatter/Gather interface. Each subwindow represents a cluster of related documents.

and request that the contents of the document be displayed for reading. Figure 7.2 presents a flow chart that captures the processing of a cluster window by a user.

Figure 7.3 presents an overview of the user-interaction process. For simplicity, the example in Figure 7.3 assumes five clusters per iteration rather than ten as depicted in Figure 7.1. Assume that a user is interested in papers written about robot planning. At the top level, the system presents the user with five cluster summaries (similar to those in Figure 7.1) that the user judges to be about "Law," "World news," "Artificial intelligence," "Computer science," and "Medicine." From these clusters, the user selects two clusters ("Artificial intelligence" and "Computer science") as being the ones likely to contain relevant papers, and he requests that the system scatter those documents into five new clusters. The user judges these new clusters to be about "Language," "Robots," "Expert systems," "Planning," and "Bayesian nets." The user then selects two clusters ("Robots" and "Planning") and requests that the system display all the titles in those clusters in the scrollable titles window. The user then scans that list of titles and picks out ones to read.

The models and simulations presented in the following sections address data collected in user studies with Scatter/Gather (Pirolli & Card, 1995; Pirolli et al.,

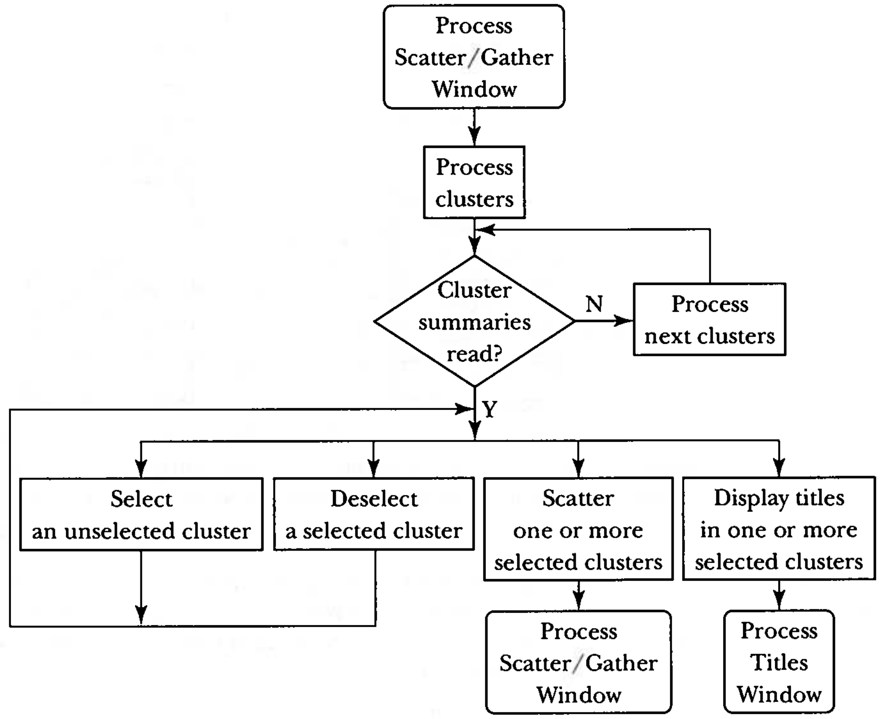

FIGURE 7.2 The task structure for using Scatter/Gather.

1996). Those studies used a document collection and a set of user tasks that had been compiled for the Text Retrieval Conference (TREC). These tasks typically asked users to find all of the documents relevant to some topic, such as new medical procedures for cancer. The tasks varied in how many relevant documents could be found.

7.4.2 Simulating Users

Pirolli (1997) developed a computer model to simulate the actions performed by individual Scatter/Gather users. The model was developed by integrating optimal-foraging models with the ACT-R architecture (Anderson & Lebiere, 2000), which is both a theory of psychology and a computer simulation environment. The resulting model was called ACT-IF (for ACT Information Forager).

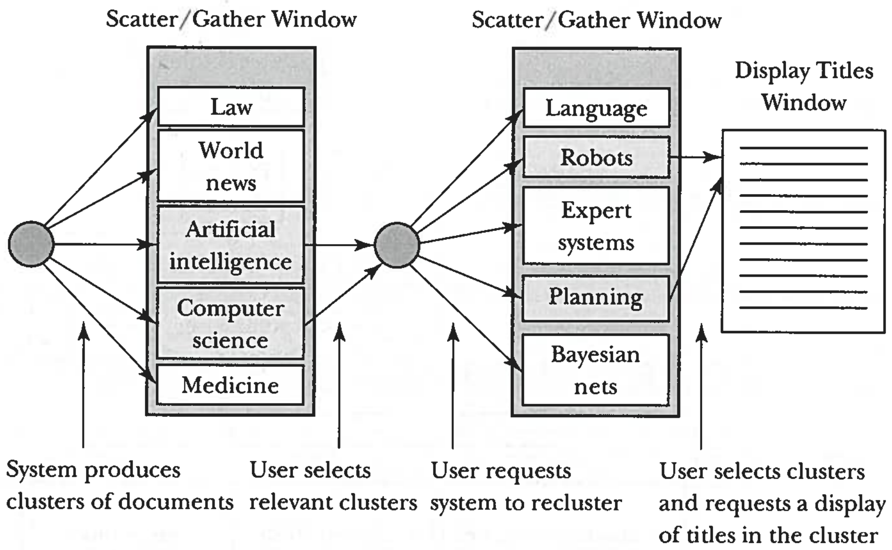

FIGURE 7.3 A schematic view of using Scatter/Gather. The user repeatedly selects (gathers) clusters from the Scatter/Gather Window and requests that the system recluster (scatter) the selected clusters until the user decides to display and scan titles from the selected clusters.

ACT-R consists of a *production memory* and a *declarative memory*. The declarative memory basically models the information being attended to and information that has been recalled from long-term declarative memory. In simulating Scatter/Gather, the declarative information mostly concerns the user's goal and the information that is presented on the Scatter/Gather screen. The production memory contains *production rules*. In the case of Scatter/Gather, the productions represent the elements of task knowledge used by the user—that is, the elements of knowledge that enable users to perform the task specified in Figure 7.2.

For instance, the following production rules represent three elements of task knowledge for working with Scatter/Gather:

P1: IF the goal is to process a Scatter/Gather Window
& there is a task query
& there is an unselected cluster
THEN select that cluster

P2: IF the goal is to process a Scatter/Gather Window
& there is a task query

& some clusters have been selected
THEN select the Display Titles Window button

P3: IF the goal is to process a Scatter/Gather Window
& there is a task query
& some clusters have been selected
THEN select the Scatter/Gather button

Each production rule has the form IF *condition* THEN *action*. The conditions of the rules are patterns that describe the cognitive state of the user. When these patterns match the cognitive state of the user, their actions may be applied to change the cognitive state and perform motor actions. Rule P1 matches when the user perceives an unselected cluster and specifies that the cluster be selected. P2 matches when the user perceives that several clusters have been selected and specifies that the user request that the system display the titles of the documents in the selected clusters. Rule P3 matches when the user perceives that several clusters have been selected and specifies that the user request that the system gather the clusters and rescatter them. In all, Pirolli (1997) represented the Scatter/Gather task using 15 production rules like P1, P2, and P3.

ACT-R operates on a *match-execute cycle*. During the match phase, the condition part of the production-rule patterns are matched against information in working memory—that is, the rules are matched against the current cognitive state of the user. Typically, more than one rule may match the cognitive state of the user. For instance, all three of the production rules listed above may match if there are selected and unselected clusters on a Scatter/Gather Window. In ACT-IF, those production rules that match are ranked by information-foraging evaluations (i.e., evaluations of the gain in valuable information that is expected by selecting a particular rule). These information-foraging evaluations assess the economics of different actions in the Scatter/Gather interface. The production rule with the highest information-foraging evaluation is selected, and then the selected rule action pattern is executed. The information-foraging evaluations are based on perceptions of information costs and value that have come to be called *information scent*.

7.4.3 Information Scent

The text in the Scatter/Gather cluster digests is much like the text used to label links on the Web. These text cues provide information scent. Users estimate the relevance of distal sources of information based on the small snippets of text available to them on their screen displays. An effective model of users' judgments of information scent is a computational model based on *spreading activation* mechanisms used in the study of human memory (Anderson, 1993; Anderson & Pirolli, 1984). Activation may be interpreted metaphorically as a kind of mental energy that drives cognitive processing. Cognitive structures, such as concepts, are called *chunks*. Activation spreads from a set of chunks that are the current focus of attention through *associations* among chunks in memory. Generally, activation-based theories of memory predict that more activated knowledge structures will receive more favorable processing. Spreading activation is the name of the process that computes activation values over a set of chunks. The spread of activation from one cognitive structure to another is determined by weighting values on the associations among chunks. These weights determine the rate of activation flow among chunks (analogous to pipes or wires with specific flow capacities). The associations and weights are built up from experience.

A specific version of spreading activation (Anderson & Lebiere, 2000) uses mechanisms based on the analysis of the requirements of an optimal memory system discussed in Section 7.3.1, on adaptationist approaches (Anderson, 1990; Anderson & Milson, 1989). This is the version of spreading activation used to develop a model of information scent. The basic idea is that a user's information goal activates a set of chunks in a user's memory, and text on the display screen activates another set of chunks. Activation spreads from these chunks to related chunks in a *spreading activation network*. The amount of activation accumulating on the goal chunks and display chunks is an indicator of their mutual relevance.

As will be discussed below, spreading activation networks can be computed *a priori* from online text collections and can be used to represent human semantic memory. Figure 7.4 presents a scenario for a spreading activation analysis. Suppose a user is looking for "information on new medical treatments and procedures for cancer" using Scatter/Gather. The representation of this goal in declarative memory is depicted by the small set of concepts linked to the concept "Information need" in Figure 7.4 (the nodes labeled "new, medical, treatments, procedures, cancer"). These nodes represent the main meaningful concepts making up the users' goal. In Figure 7.4, the user is looking at a browser that has retrieved a set of documents. One of the cluster-digest summaries is the text "cell, patient, dose, beam, cancer." The representation of this part of the browser display is depicted by the network of nodes linked to "Screen text" in Figure 7.4. The figure also shows that there are links between the goal concepts and the text-summary concepts. These are associations between words that come from past experience. The associations reflect the fact that these words co-occur in the

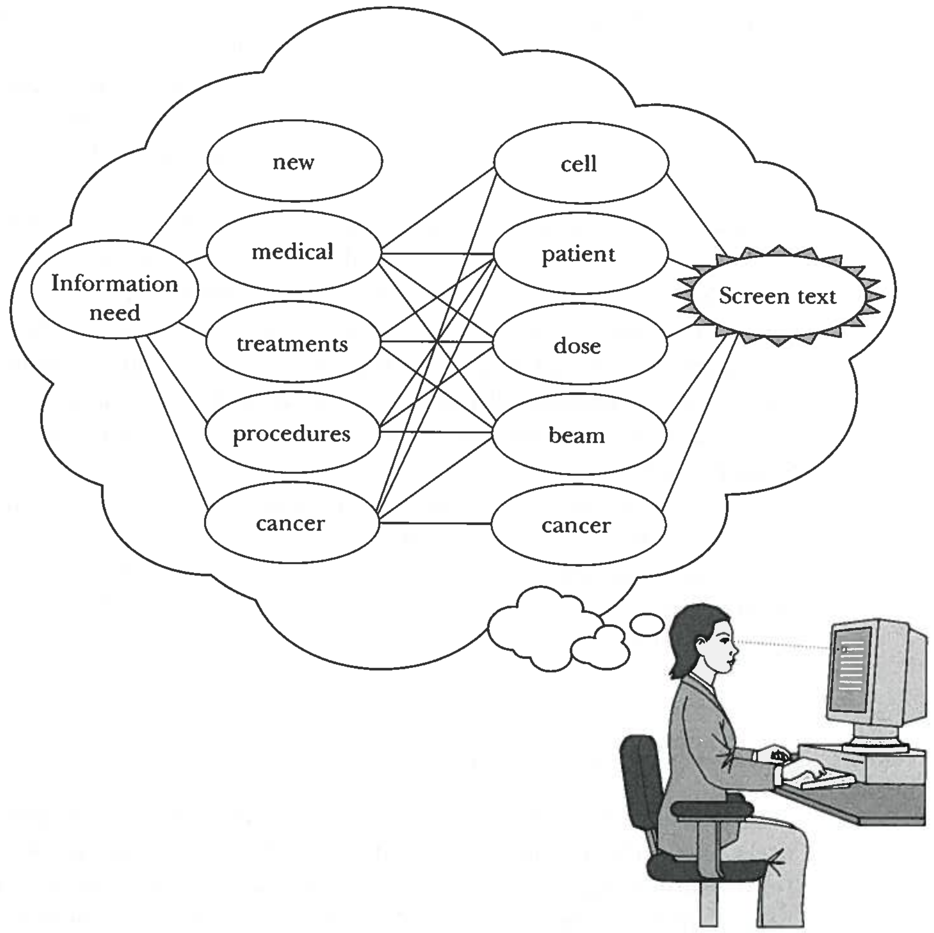

FIGURE 7.4 Information scent.

users' linguistic environment. For instance, the words "medical" and "patient" co-occur quite frequently, and they would have a high weighting of inter-association. Spreading activation would flow from the goal, which is the focus of attention, through the interword associations, to words in the text summary. The stronger the associations (higher weights or strengths that reflect higher rates of co-occurrence) the greater the amount of activation flow. If the goal and browser text are strongly associated, then we expect people to judge them as being highly relevant to one another. At least implicitly, this is what the interface designers of browsers are trying to do when they select small text snippets to communicate the content of large documents to users. They are trying to pick words that people will judge as relevant to their queries. Spreading activation may be used to predict these memory-based judgments of reminding and relevance that are key components of surfing the Web.

Some Web-usability guidelines (User Interface Engineering, 1999) center on the design of effective links that provide "good" information scent that draw users to the information they seek. It would be useful to have automated methods that predict the effectiveness of the information scent of links with respect to user goals. Pirolli and Card (1999) showed how an information-scent model could make accurate predictions of user judgments of information scent (Figure 7.5), and they then also showed how the information scent provided by the Scatter/Gather interface accurately reflected the amount of relevant information in clusters described by the text in cluster digests. Together, these analyses suggest that the text on the cluster digests of the Scatter/Gather browser accurately communicated the location of relevant information. Information-scent models might provide automated methods for evaluating the content design of links on Web pages.

7.4.4 Information-Foraging Evaluations

The information-scent model also provides the basis for understanding how users assess the economics of different actions available in Scatter/Gather. Information-foraging theory assumes that users' choices are directed by cost-benefit analyses: the costs (usually in terms of time) of choosing a particular course of action weighed against the value to be gained by that action. In the case of Scatter/Gather, using the TREC tasks, the value was the number of relevant documents identified by a user. The basic costs of different user and system actions could be determined, in part, by using estimates from the literature (Olson & Olson, 1990), or by measuring system or user times on just the component actions themselves. Figure 7.6 presents a schematic of the action choices that are captured by productions P1, P2, and P3. The square objects represent states of Scatter/Gather interaction (interactions with Scatter/Gather windows are at the top and those with Display Titles windows are at the bottom), and the arrows represent transitions among interaction states that result from the execution of a production rule. For current purposes, let us represent the information-foraging evaluation of productions by the notation *Eval[P1: choose cluster]*, which should be taken to mean, the evaluation of production P1 that executes

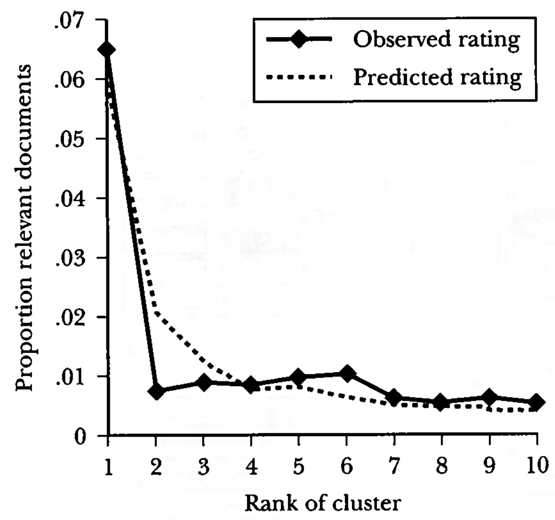

FIGURE 7.5 **Ratings of cluster summaries predicted by information scent plotted against the observed ratings produced by users.**

the action of choosing a cluster. The following sections describe the evaluations for Productions P1, P2, and P3.[15]

Cluster Selection

The analysis of cluster selection can be developed by analogy to models of optimal-diet selection in optimal-foraging theory (Stephens & Krebs, 1986). The Scatter/Gather display (Figure 7.1) presents users with the number of documents in each cluster, and this can be used by users to estimate the costs of finding relevant documents in each cluster. Assume that the user decided to go through every document title in a cluster and then saved the relevant documents for later reading.[16] There would some cost associated with evaluating every document, and some additional cost of saving the relevant documents. So the profitability for a cluster could be characterized by the following,

[15] In Pirolli and Card (1999), a notation was used that was more consistent with the optimal-foraging literature: π = *Eval[P1: choose cluster]*, R_{SG} = *Eval[P2: rescatter]*, and R_D = *Eval[P3: display titles]*.

[16] This could be achieved by displaying the cluster, scanning and scrolling the displayed list, and cut-and-pasting the relevant titles to a file.

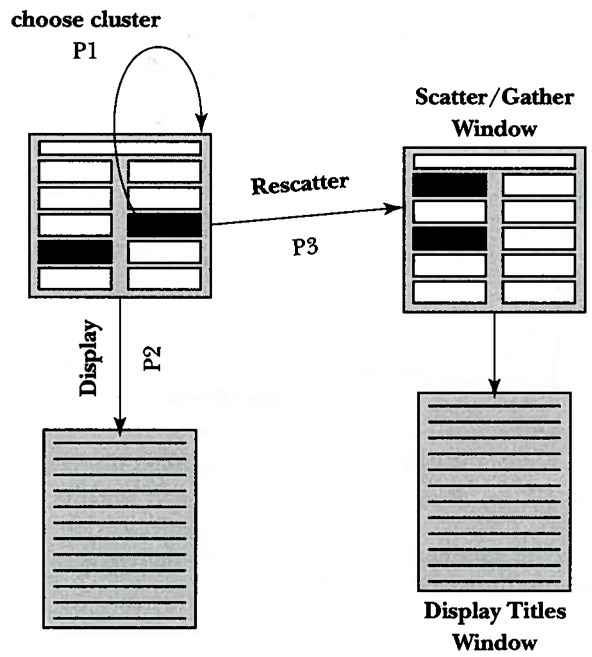

FIGURE 7.6 Basic moves and evaluation parameters for the Scatter/Gather task.

S = proportion of relevant documents in a cluster according to information scent,

N = total number of documents in a cluster (from the Scatter/Gather display),

V = the number of valuable (relevant) documents in a cluster (judged by information scent)

$= N \times S$

C_1 = the time cost of processing each document,

C_2 = the time cost of processing a relevant document,

Eval[*P1: choose cluster*] = profitability of a cluster in terms of number of valuable documents per unit time,

$= V / [(N \times C_1) + (V \times C_2)]$

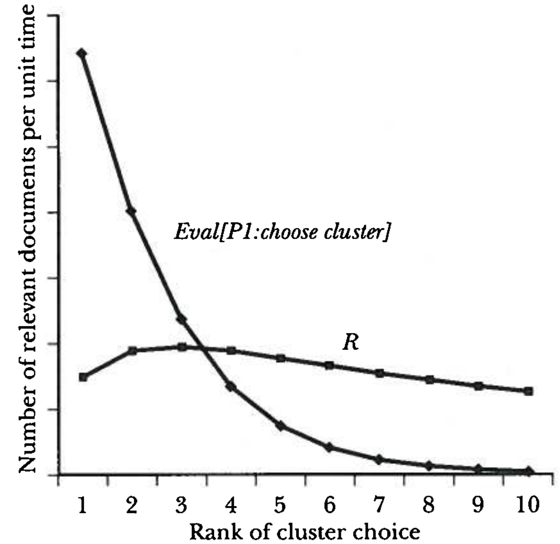

FIGURE 7.7 An optimal foraging model applied to Scatter/Gather. Clusters are ranked by their profitability in this graph. R is the rate of gain for the collection of clusters of rank 1 through i. Note that these curves vary for every Scatter/Gather Window.

Figure 7.7 shows a typical distribution of profitability values, *Eval*[*P1: choose clusters*], over clusters. Suppose the user ranked the clusters by their profitability (which is determined by the proportion of relevant documents they contain), and then selected them in rank order up to some threshold (which is the optimal strategy). Adding more clusters to the "information diet" affects the overall expected rate of gain, R = relevant documents/time, as shown in Figure 7.7. In ACT-IF, R corresponds to the maximum value of the evaluations of P2 or P3,

$$R = \text{MAXIMUM}(\textit{Eval}[\textit{P2:display titles}], \textit{Eval}[\textit{P3: rescatter}]).$$

The evaluations of P2 and P3 are described in more detail below. In other words, ACT-IF will continue to choose clusters (using production P1) until either P2 or P3 has a higher evaluation. At that point, the expected rate of return (R) is maximized. Choosing more clusters will diminish the rate of gain of valuable documents per unit time.

Adding the topmost-ranked clusters tends to improve R because they contain proportionally more relevant documents. Adding the bottommost-ranked clusters tends to decrease R because they contain proportionally more irrelevant documents, which are a time cost for users. The optimal set of clusters to select is the one that produces the highest value of R. When the curves for *Eval[P1: choose clusters]* and R are drawn as in Figure 7.7, this optimal point occurs just to the left of the point at which the curves cross. Empirical studies (Pirolli & Card, 1999) indicate that users select clusters in accordance with this model. Users select an optimal-information diet from each Scatter/Gather screen.

Scanning Displayed Titles

Eval[P2: display titles] is the rate of gaining relevant documents from selected clusters. At some point, the user selects a set of clusters, displays the titles, and scans them for relevant documents. The document titles are unordered in the scrollable display window. Consequently, the total number of relevant documents encountered as a function of user scan time will be linear. The rate of these encounters will depend on the proportion of relevant documents in the clusters that were selected to be displayed. In general, the proportion of relevant documents in clusters presented to users depends on how many times the user has gathered and rescattered clusters. As the user invests more time in gathering and scattering clusters before displaying titles, the proportion of relevant documents increases in the available clusters, and the rate at which the user will encounter relevant documents when document titles are displayed will also increase. From the proportion of relevant documents in selected clusters, one can determine the rate at which relevant documents will be encountered by a user who chooses to display the titles in those clusters.

Scattering and Gathering Clusters vs. Scanning Displayed Titles

If an ideal user chooses to repeatedly scatter and gather clusters in an optimal fashion, then the total number of documents in the selected clusters will decrease, but the number of relevant documents will decrease at a slower rate. The proportion of relevant documents under consideration will improve as a function of user time. This will improve the rate at which users gather relevant documents once they decide to display clusters. From this analysis, one can determine *Eval[P3: rescatter]*, which is the rate of gain of relevant documents that would be produced by one more round of gathering and scattering clusters. Figure 7.8 shows a plot of *Eval[P2: display titles]* and *Eval[P3: rescatter]* over time for an

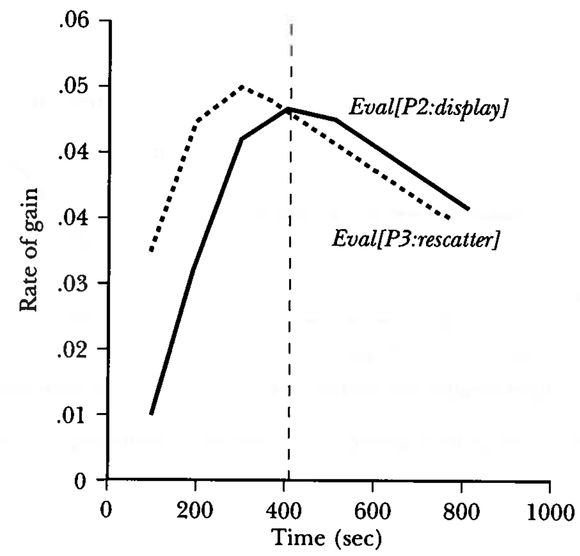

FIGURE 7.8 Rate of gain evaluations for gathering and rescattering clusters *(Eval[P3: rescatter)* vs. displaying documents *(Eval[P2: display])*, as a function of task time.

optimal user interacting with Scatter/Gather. Early in the task, it is optimal to continue scattering and gathering clusters, but there comes a point at which it is optimal to display the titles.

7.4.5 Simulating Users and Evaluating Alternative Scatter/Gather Designs

An engineering model derived from the ACT-IF model of Scatter/Gather users was used to evaluate hypothetical changes to the Scatter/Gather system. That is, rather than perform costly testing of real users, the designs were tested against simulated users. Figure 7.9 shows predictions for two potential design improvements: an increased speed of interaction (by improving the speed of gathering and rescattering clusters) and an improved clustering that puts more relevant documents into fewer clusters. To the question "which improvement is better," the answer is "it depends on the task." If the user has lots of time (soft deadline) then improved clustering is better, but if the user has little time (hard deadline)

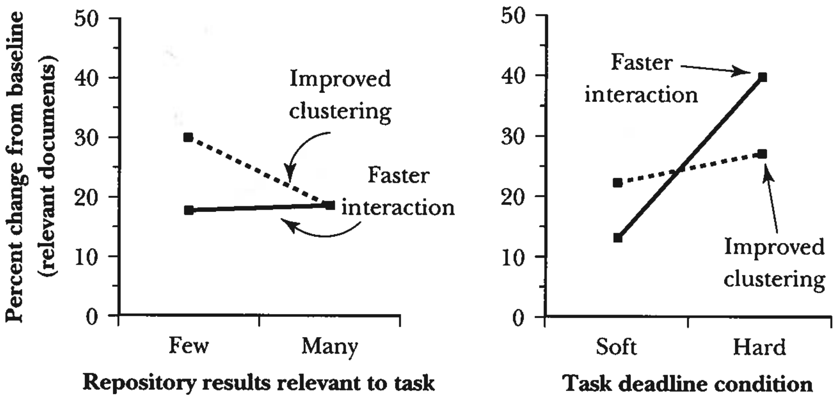

FIGURE 7.9 Predicted tradeoffs of two potential design improvements to Scatter/Gather.

then faster interaction is better. These kinds of "it depends" tradeoffs are rampant in interface design, but user simulations can be more precise about their nature.

7.5 CASE STUDY: THE WORLD WIDE WEB

Users of the Web work in a much richer environment than the Scatter/Gather system, and they often perform far more complicated tasks than simply finding relevant documents. Ongoing research aimed at extending information-foraging theory to understanding Web users and Web designs provides a useful framework that provides much-needed coherence and insight. For instance, the basic notions that users follow information scent and that they optimize the economics of their information seeking helps in understanding behavior at the level of individuals as well as large-scale aggregates of users (such as the behavior of all the visitors to a Web site). These notions also lead to usability guidelines, new user interfaces, and technologies for performing automated analyses of Web site designs.

7.5.1 Information Scent as a Major Determinant of Web User Behavior

Effects at the Level of Individual Users

As was the case with Scatter/Gather, the structure of Web tasks depends in part on the organization of content and the kinds of operations that the user can perform in browsing and searching. Figure 7.6 provided a schematic representation of the Scatter/Gather task structure by identifying common states of interaction and common moves among states. Figure 7.6 is an informal example of a *state-space diagram,* which is a common device used in the cognitive sciences for visualizing problem-solving activity and other forms of complex cognitive activity. A similar state-space representation has been developed for Web interaction. Figure 7.10 and Plate 5 of the color insert present two Web Behavior Graphs (WBGs) depicting the behavior of a user performing two Web tasks: (a) finding a specific movie poster (top of Figure 7.10) or (b) finding a scheduled performance date for a specific comedy troupe (bottom of Figure 7.10).[17] A WBG is a kind of state-space diagram and, more specifically, it is based on *Problem Behavior Graphs,* which can be found in Newell and Simon's (1972) classic investigation of human problem solving. Each box in the WBG represents a state of the user and system. Each arrow depicts a move from one state to a new state. For instance, the sequence S2 → S3 in Figure 7.10 represents that a user visited a particular Web page (S2) and clicked on a link and then visited a second Web page (S3). Double vertical arrows indicate the return to a previous state, augmented by the user's experience of having explored the consequences of some possible moves. Time in the diagram proceeds left to right and top to bottom. Moves within the same Web site are surrounded by the same color in the WBG. This reveals moves within and across Web sites.

Information scent has a significant impact on Web users. The links selected by users tend to be the ones that are scored as having the highest information scent. Poor information scent causes users to flounder because they tend to select links that lead them to useless parts of the Web. This can be seen by inspecting Figure 7.10. The WBG at the top of the figure is much bushier than the WBG at the bottom, involving many more returns to pages previously visited. This

[17] The WBGs presented in this chapter are simplified versions of those presented in published reports (Card et al., 2001).

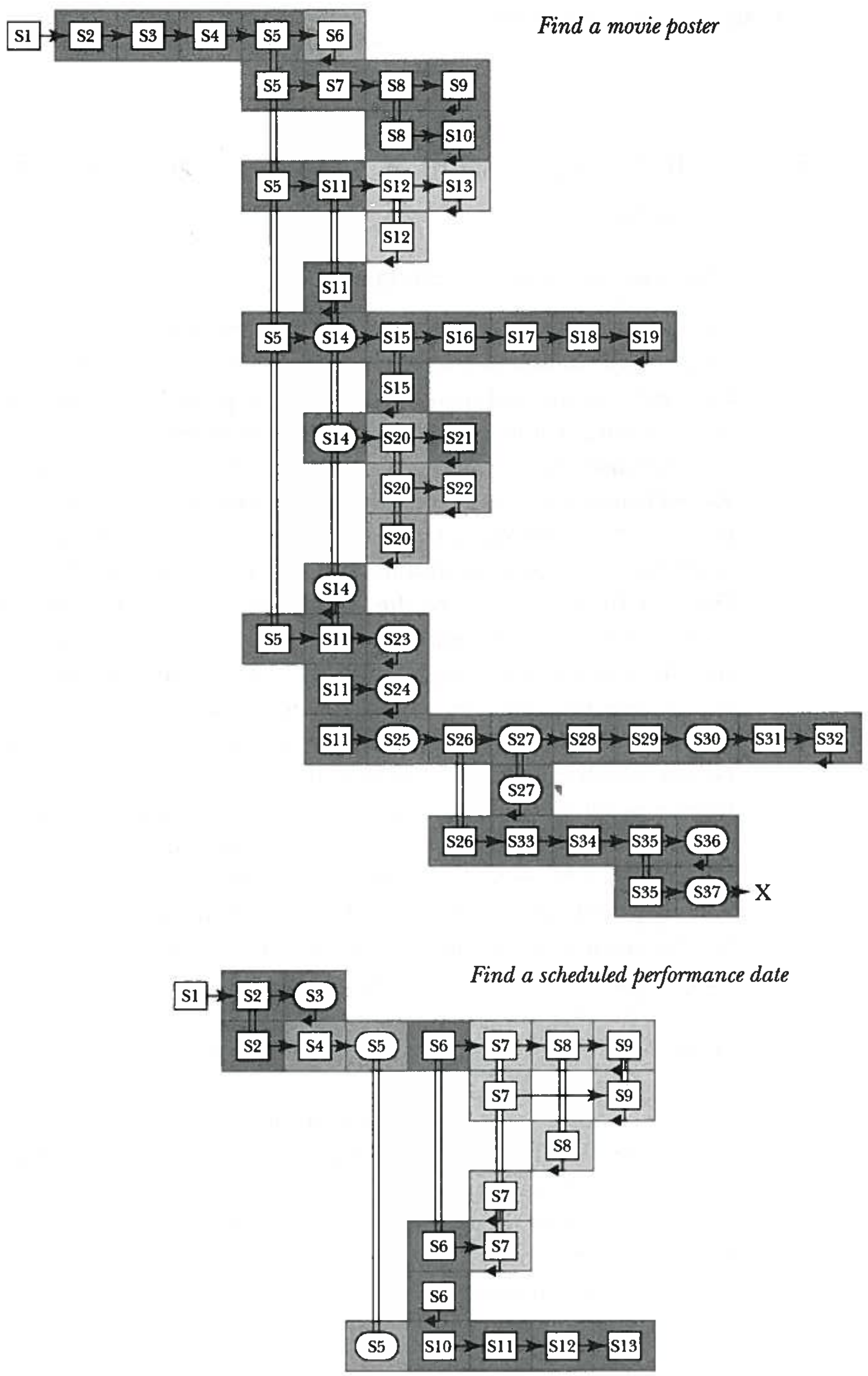

FIGURE 7.10 Web Behavior Graphs for a user on two tasks. Each box represents a state of the user-system interaction. Each arrow represents a state change (move). Interaction proceeds left to right, top to bottom. Each move to the right is a move to a new state. Moves down and to the right represent a return to a previous state. Background colors surround interactions with the same Web site. (See Plate 5 of the color insert for a color version of this figure.)

bushiness, or high *branching factor,* in the top WBG is an indication that the user had difficulty finding information scent that led to useful information. Often, users appear to switch from one Web site to another, or they switch from browsing to searching, when the information scent gets low. The propensity to switch to another information patch (e.g., another site or a search engine) when the information scent becomes low is consistent with optimal-foraging theory (McNamara, 1982).

Effects at the Aggregate Level

At the individual level, highly informative and accurate information draws users directly to their goal because they make accurate decisions at every state. The following analysis shows that the accuracy of these scent-based decisions has a profound overall qualitative impact on navigating through information structures like the Web. Concretely, designers can use a number of methods to make link descriptors have more accurate information scent. For instance, including more summary words (up to about 20) seems to improve information scent (User Interface Engineering, 1999). Generally (but not for all tasks), text appears to be superior to images (thumbnails). A combination of text and words called *enhanced thumbnails* (Woodruff, Rosenholtz, Morrison, Faulring, & Pirolli, 2002) combines the search benefits of image thumbnail summaries and text summaries for links (Figure 7.11 and Plate 6 of the color insert).

How important is it to ensure that users get accurate information scent about links and make accurate decisions about navigation? In a study of text, image thumbnails, and enhanced thumbnails as link summaries, Woodruff et al. (2002) calculated the *false-alarm rates* for the different link summaries. These false-alarm rates were the probabilities that a user would incorrectly follow a link—that is, the user thought that the link would lead to relevant information but it did not. These false-alarm rates were in the range of .01 to .15. Although the absolute sizes of these false-alarm rates seem small, such variations can have a dramatic impact. This can be illustrated by considering an idealized case of searching for information by surfing along links in a Web site.[18] Assume that the imaginary Web site is arranged as a tree structure with an average branching factor b. Assume that a user starts at the root page and is seeking a target page that is depth d from the root. If the false-alarm rate, f, is perfect, $f = 0$, then the user will visit d pages. This cost grows linearly with d, the distance of the target from the root. If the false-alarm rate is maximum, $f = 1$, then the user will visit half the

[18] The following analysis is based on a method for the analysis of heuristic search costs developed by Huberman and Hogg (1987).

FIGURE 7.11 Enhanced thumbnails. (See Plate 6 of the color insert for a color version of this figure.)

pages in the Web site, on average. This cost grows exponentially with *d*, since the number of pages grows exponentially with depth.

Figure 7.12 shows the effects of perturbations in false-alarm rates more concretely. The figure displays search-cost functions for a hypothetical Web site with branching factor $b = 10$. The curves represent cost functions for links with false-alarm rates of $f = .100$, .125, and .150. One can see that the search-cost regime changes dramatically as f becomes greater than .100. Indeed, for a branching factor of $b = 10$, there is a change from a linear search cost to exponential search cost at the critical value of $f = .100$. Small improvements in the false-alarm rates

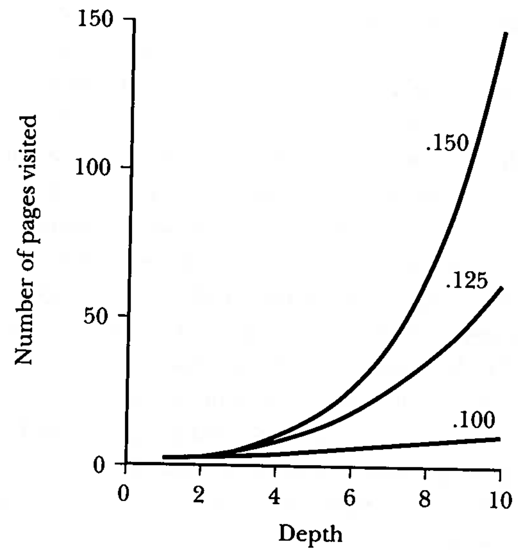

FIGURE 7.12 Changes in the search-cost regimes with changes in information scent. Search cost is in terms of number of Web pages visited in a hypothetical Web site that is 10 levels deep with 10 branches per level. Quantities labeling the curves indicate the probability of a false alarm.

associated with individual links can have dramatic qualitative effects on surfing large hypertext collections. In other words, there is a qualitative shift in the cost structure of navigation costs due to changes in the accuracy of information scent. Web usability experts (User Interface Engineering, 1999) have good reason to recommend that designers focus on designing Web sites with good information scent.

One business model that emerged for Web site producers was to sell advertising and to try to keep users on their sites for as long as possible. Sites that did this effectively were called *sticky*. As discussed earlier, information scent (or the lack of it) appears to determine when users leave a site. A model has been developed (Huberman, Pirolli, Pitkow, & Lukose, 1998) that predicts the distribution of moves (i.e., page visits) that users will make at a Web site. The underlying model is a kind of *random walk model* that has applications ranging from the study of Brownian motion in physics to real options in financial economics. Brownian motion refers to the trajectories made by particles such as pollen suspended in water. To model these trajectories, it is useful to consider each particle as making a random walk in space—each successive step will be a random move in some direction, but the next location must be somewhere near the previous location. Sometimes this model is elaborated to include a notion that there is drift that biases the random moves in some direction. A question (first analyzed by Albert Einstein) is how long it takes a particle to cross an imaginary threshold. The probability distribution of the time it takes particles to cross a threshold (called the *first passage* time) is called an *Inverse Gaussian* distribution. The analogy in the case of the Web is that users correspond to particles moving in a state space of Web pages. Pages are characterized by their information scent (instead of spatial locations), and the user makes moves that drift towards higher scent pages but with some degree of uncertainty (randomness). It is assumed that pages that are linked together tend to be similar in information scent (or at least more similar than pages separated by many links). Users move through this state space until they hit a threshold—for instance, they encounter a page that has high relevance or the information scent gets too low and they switch to another site. If this random walk model characterizes Web users, then we should expect to see an Inverse Gaussian distribution in the time users spend at a Web site (before they decide they have found something or they need to stop and go do something else).

This Inverse Gaussian distribution is shown in Figure 7.13 where it has been fit to usage data from a Web site at Georgia Tech in 1994. The mean and variance of this distribution—that is, the mean number of moves made at a Web site and their variance—is determined by the users' information scent thresholds as well as statistical properties that describe the relationship between the value of the currently visited page and the expected value of a linked page. The statistical properties of the variation of information scent over a Web site determines the particular Inverse Gaussian distribution of moves for a Web site. This is a rather precise way to say that information scent determines the stickiness of a site.

7.5.2 Simulated Users and Usability Evaluation

Chi, Pirolli, Chen, and Pitkow (2001) recently developed techniques aimed at (a) predicting usage patterns at Web sites as well as (b) inferring user goals from existing logs of user behavior. Although these simulations are not as detailed as the ACT-IF model of Scatter/Gather, they can become the basis for engineering tools for Web site designers. The Web usage simulation uses a Web crawler to analyze the content and linkage structure of a Web site. For a given user need (represented by a set of query words), the system computes the link scent for every link on every page. From this, the simulation calculates the probability that a user will go down each link emanating from a page. The simulation proceeds by

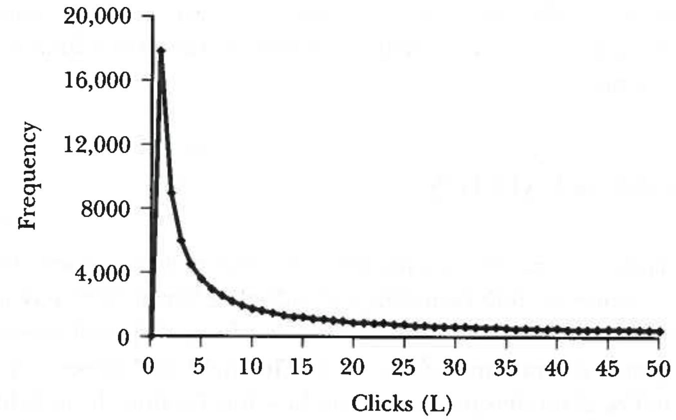

FIGURE 7.13 A fit of the Inverse Gaussian distribution (solid line) to the frequency of users who click L times at the Georgia Tech Web site (data from 1994).

flowing simulated users through the Web site with the flow among pages determined by information scent.

Chi and colleagues also developed a method for inferring the information goal of a user based on the user's traversal path through a Web site. The technique is based on the observation that users communicate a great deal of information about their goals with every link choice they make. At any point in a user's traversal through the Web site, that user has expressed interest in certain links rather than others. The technique developed by Chi and colleagues takes the documents visited by a user and infers a set of keywords that best represent the path taken by the user. Studies by Chi and colleagues indicated that these keywords were effective in communicating the path taken by the user.

A similar system called Comprehension-based Linked Model of Deliberate Search (CoLiDeS) (Blackmon, Polson, Kitajima, & Lewis, 2002) has been developed to evaluate Web sites. CoLiDeS uses a model of information scent based on LSA (Landauer & Dumais, 1997), rather than spreading activation, and achieved success in predicting unfamiliar and confusable links. CoLiDeS is based on an earlier modeling system called Linked Model of Comprehension-based Action Planning and Instruction (LICAI) (Kitajima & Polson, 1997) that addressed how users learn by exploring computer applications by following the labels on interface items such as menus. Such label following is a common strategy in performing actions in computer applications, mobile devices, and control panels. At each choice point in a menu or submenu, the user must assess the information scent of the labels with respect to their goal. Consequently, user models that incorporate predictions about information scent should be applicable not only to the Web, but also to mobile devices and consumer products with menu-driven control panels.

7.6 CURRENT STATUS

This chapter has necessarily focused very narrowly on information-foraging theory as it relates to understanding individual behavior. There is also an emerging field of work that examines the complex emergent phenomena that arise from many agents interacting on the Web. This field of *Internet ecology* uses complex computational models to understand how local actions by individuals are related to complex global phenomena, both social and computational in nature. The relationship between the surfing of individuals and the Inverse Gaussian distribution of number of clicks at a site was one example of this kind of research. This area of research frequently draws upon nonlinear dynamics and statistical physics (recall the application of models of Brownian motion to predict the distribution in Figure 7.13). Some research in this field includes:

- Internet congestion (Huberman & Lukose, 1997). The Internet is a public good. For the most part, users are not charged for its use. Users may rationally decide to consume bandwidth in a greedy manner while thinking that their behavior will have little effect on the Internet as a whole. This leads to a degradation of Internet performance. A specific model of this "social dilemma" leads to accurate predictions of the statistics of short-lived "spikes" of Internet congestion.
- The power law distribution of users over sites (Adamic & Huberman, 2000). The number of visits received by Web sites (whether within a particular category, or across all) follows a power law. This is also a characteristic of "winner-take-all" markets. Dynamical models of Web site popularity predict this distribution.

The Internet and the Web provide a fascinating laboratory for detailed quantitative measurements of large-scale ecological phenomena. Now that this laboratory is available, we should expect to see more theories that relate individual behavior to large-scale phenomena.

One novel aspect of information-foraging theory is that it begins to deal with the interaction of the user with information content, rather than just application interfaces. There has been very little research on understanding how people interact with *rich content*—that is, content other than text. Although there has been considerable research in psychology and artificial intelligence on image understanding (for instance, line detection, object recognition), there has been no detailed psychological research on how people browse or navigate through hypermedia that includes images, video, animation, and so forth. There is also a gap in our understanding of how users control their visual browsing of individual displays of content. There has been very little in the way of detailed research on how people visually scan and search for information on Web pages, and such theories could be of great use to designers interested in visual layout.

The studies that have been discussed in this chapter mainly focused on relatively well-defined information-seeking tasks. Most tasks on the Web include information seeking, but are typically broader or more ill defined in nature. They include tasks such as trying to make sense of medical publications about a disease affecting a loved one, choosing a career, finding a good graduate school, and so on. These ill-structured problems dominate our everyday life (Reitman, 1965; Simon, 1973), and they are the ones that require the user to seek and understand as much information as feasible. Information-foraging theory and, more broadly, HCI, will increasingly need to address these complex ill-structured problems and how they might be aided by enhancements in information technology.

AUTHOR NOTES

The writing of this chapter was supported, in part, by Office of Naval Research Contract No. N00014–96–0097. I would like to thank Thomas Green, Julie Bauer Morrison, Mark Perry, and Pamela Schraedley for their reviews. I am also indebted to the students of Virginia Tech's CS 5724 class: Ndiwalana Ali, Stephen Foret, Christian Schaupp, and Akshay Sivananda, who provided thoughtful reviews.