Chapter 2. Information Seekers and Electronic Environments

Throughout our lives we develop knowledge, skills, and attitudes that allow us to seek and use information. This chapter introduces the notion of personal information infrastructure, which will be used to describe this complex of knowledge, skills, and attitudes. It also introduces the notion of interactivity, a key characteristic of computer technology that allows information seekers to use electronic environments in ways that emulate interactions with human sources of information. The chapter also provides an overview of the technological developments that underlie information seeking in electronic environments.

PERSONAL INFORMATION INFRASTRUCTURES

The primary activities of scientists, physicians, businesspersons, and other professionals are devoted to gathering information from the world, mentally integrating that information with their own knowledge--thus creating new knowledge--and acting on this new knowledge to accomplish their goals. Most often, this knowledge and the consequences of using it are articulated to the external world as information. All humans develop mental structures and skills for conducting such activities according to their individual abilities, experiences, and physical resources. An individual's collection of abilities, experience, and resources to gather, use, and communicate information are referred to as a *personal information infrastructure*. A personal information infrastructure is a collection of interacting mental models for specific information systems[1]; mental models for events, experiences, and domains of knowledge; general cognitive skills (e.g., inferencing, recognizing salience) and specific cognitive skills related to organizing and accessing information (e.g., filing rules, reading); material resources such as information systems, money and time; metacognitive resources for planning and monitoring thought and action; and attitudes toward information seeking and knowledge acquisition. Figure 2.1 illustrates the main components of a personal information infrastructure. The level of development of an individual's information infrastructure is roughly analogous to a level of information literacy. Note that personal information infrastructure as used here is much broader than the way information infrastructure is used in the electronic networking literature where the focus is on the physical and computational resources (e.g., Kahin, 1992).

[Insert Figure 2.1 about here]

Cognitive psychology has posited many levels of cognitive processes to explain human mental activity. Theories of memory, decision making, and problem solving include explanations of how knowledge is represented at different levels of granularity. For example, cognitive scientists offer theories for short term and long term memories (Estes, 1982; Wickens, 1987), semantic networks (Quillian, 1968), frames (Minsky, 1975), scripts (Schank & Abelson, 1977), and mental models (Johnson-Laird, 1983). *Mental models* are dynamic mental representations of the real world (Johnson-Laird, 1983;

Norman, 1983). People construct, then draw upon mental models to predict the effects of contemplated actions, i.e., they make inferences based on 'running' particular mental models. Information seekers develop and use mental models for a variety of mental and physical objects, including information objects and different domains of knowledge. A mental model for a particular information object such as a book allows one to base expectations about how to begin and proceed in reading and to estimate how much effort will be required. A mental model for the domain (topic area) related to the book's content allows the reader to integrate information (understand) as reading progresses. Mental models account for expectations and therefore learning and change in behavior. A personal information infrastructure includes the various mental models an individual has developed for different information systems and domains of knowledge.

Experience with a variety of information problems and systems leads people to develop general knowledge about how information is organized, and skills for facilitating access. We learn to recognize the advantages and limitations of general organizational structures such as lists, arrays, hierarchies, and networks and how to leverage the advantages and mitigate the limitations. At the most basic levels, skills include memory processes such as rehearsal, association, and chunking (Lindsay & Norman, 1977) and strategies such as use of mnemonics. At more formal levels, they include the strategies and relationships we learn and develop throughout our lives. Greeno (1989) distinguishes symbolic/abstract knowledge from mental models to account for these types of knowledge. For example, we learn filing rules such as alphabetical, chronological, or positional orderings that facilitate subsequent retrieval. These rules are generic and serve as defaults for orderings when we encounter a new domain. Particular domains (e.g., biological classes, library classifications, chemical structures) use specific orderings that must be integrated in mental models for the domain. As we will see in Chapter 4, recognizing the distinctions and overlaps between generic and specialized knowledge is one characteristic we call expertise in a domain. Well-developed personal information infrastructures allow people to look for organizational rules in new domains before applying default rules. The organizational rules become defaults for experts and thus organizations in our personal lives reflect the domain organizations in our professional lives.

We also learn about information sources such as books, journals, encyclopedias and indexes. Our general knowledge about what is typical about these generic type of systems overlaps with our particular mental models for those systems. Furthermore, the formal strategies for using particular information sources (e.g., use of a back of the book index) that are part of our mental models for those systems are generalized and serve as the basis for heuristic or analogical strategies when we encounter new systems. We develop knowledge about areas of personal interest and relationships with others with expertise in those areas so we can exchange information when needed. As we gain experience with information problems, we strengthen our general information-seeking knowledge and skills just as we develop our knowledge and skills in other general cognitive processes such as listening, reading, writing, speaking, reasoning, and decision-making.

Many cognitive theories include an executive process that controls and monitors the various perception, memory, computation, and motor processes. A popular view of this executive process is termed metacognition (e.g., Flavell, 1985). Metacognitive activity refers to our ability to reflect on our own thoughts and actions in the past, monitor them as they proceed, and plan which ones to take to meet our needs. Our personal information infrastructure is guided by metacognitive activity directed at meeting situated information needs. Metacognition determines that we need information, enables our general information-seeking knowledge and our mental models for systems and domains, and monitors progress. Metacognition is influenced by affective states such as motivation and attitude and by physical states such as fatigue and comfort levels.

Material resources that make up the personal information infrastructure include: people, books, computers, telecommunications lines, and all the other tangible things we use to gather, generate, manage, and communicate information. Material resources also include money and time we have available to use and maintain these resources. These physical components of our personal information infrastructures are most readily affected by sociological and technological developments. To augment our memories, we accumulate huge collections of paper covered with relatively permanent, visually accessible symbols and marks. To organize these collections of paper,we acquire drawers, cabinets, shelves, libraries, and archives. To replicate and distribute items from these collections we use copiers, mail and courier services, and telefacimile. To acquire new information, we maintain personal reference collections, hire clerical support staff, nurture networks of colleagues, contract with research companies, and visit libraries. All these objects, people, communication channels, and strategies are part of a personal information infrastructure that individuals develop to accomplish their goals. Individuals support many layers of personal information infrastructure to serve long and near-term goals and many intermediate goals within the infrastructure itself.

Influence of Electronic Digital Technology on Personal Information Infrastructures. Electronic technology affects personal information infrastructures at all levels. Most obviously, technology affects the material resources of our personal information infrastructure by presenting new objects (e.g., computers, disks) to purchase and manipulate. To acquire new information, we use online databases, electronic bulletin boards, and local magnetic or optical databases. Increasingly, sources of information are made available only in electronic form. The physical changes electronic technology bring are highly dependent on material wealth and moderately dependent on individual ability.

Electronic environments are also affecting the cognitive and affective components of our personal information infrastructures. Long term memory is augmented by magnetic media and digital signals; magnetic, optical disks and backup systems store information organized in files, databases, and hyperdocuments; and information is copied and shared through local and global communications networks. We must develop new mental models for different systems so that we are able to use them and develop new experience bases that allow us to apply them appropriately. Technology augments our cognitive skills in several ways: by providing online assistance in selecting and using information sources (e.g., context-sensitive help, online reference manuals, spelling and grammar checkers, thesauri and encyclopedias, cut-and-paste tools, etc.); by broadening the proximity of personal networks (through electronic mail and bulletin boards); by extending our personal knowledge; and by changing the strategies we use for seeking and acquiring information (e.g., browsing, string search, relevance feedback).

Adding electronic technology to our information infrastructures can have significant impacts on our cognitive activity. Electronic technology in general can amplify and augment our abilities and performance (Engelbart, 1963) as well as disorient and confuse us (Mantei, 1982). Computer applications such as electronic text are causing changes in the fundamental processes of writing and reading (Bolter, 1991), and electronic spreadsheets have enabled iterative decision making for individuals and businesses. High performance computing and scientific visualization techniques are facilitating new insights into complex scientific, medical, and engineering problems, enabling new discoveries and speeding progress. Calculators and computers have changed the K-12 mathematics curriculum by legitimizing systematic guessing--students are now taught iterative "guess and check" strategies for solving problems as well as formal analytic strategies.

In addition to amplifying and augmenting our cognitive activity, electronic technology affects our metacognitive activity by changing our expectations. Using a word processor with spell checking, we expect the system to find spelling errors and type with abandon; likewise we have come to expect laser-quality output. Using database systems for managing lists leads us to expect to be able to display information in a variety of sorted forms with the press of a key, and we are disappointed when we look at a printed list that is not ordered in a manner that is optimal for our immediate needs. We expect rapid response and comprehensive scope when searching for information. We come to expect to use string search capabilities to locate words or phrases and often incorrectly expect that string search will locate concepts or ideas rather than literal words or phrases. Novices often expect that information obtained from a computer will be more exhaustive in scope and more accurate. Such expectations change the actions we take as we seek information as well as the way we create and use information. These changes are viewed as cognitive laziness or sloppiness by some, and as mental emancipations by others.

The material resources that comprise personal information infrastructures maintained by professionals today are physical and virtual, static and dynamic, proximate and global. In fact, managing the material resources has become complex and expensive, requiring unique skills and technology. Increasingly, the material resources are managed and used from computer workstations connected to global networks rather than through personal contacts and physical manipulation of objects. Since we have limited short-term memory resources, any cognitive resources devoted to managing the personal information infrastructure itself are not available to the information problem at hand. Since the information problems we face are increasingly complex, they demand maximum perceptual and cognitive resources. We can ill-afford to devote significant mental resources to managing the material resources; we require easily manipulable interfaces.

Electronic augmentations of our personal information infrastructures can make us more productive and thus more emotionally fulfilled and happy. Moreover, these environments offer the freedom to explore new areas of interest easily and rapidly. However, the complexity and ever-changing nature of the tools can cause stress and anxiety as well. Furthermore, we can come to feel dependent on the technology, requiring computers to access our written notes, files, and mail. These environments can also cause alienation if we have fewer and fewer personal interactions with people having common interests.

A key feature of living in an information society is managing the many potential resources technology has enabled. We find ourselves interacting with complex, electronic environments to generate, store, manipulate, access, and use information resources. Electronically-augmented personal information infrastructures affect us physically, cognitively, and emotionally. Physically, they cause us to use fewer large motor movements and more small motor movements; to be more sedentary--sitting, typing, reading from screens, subvocalizing--rather than actively moving from place to place, manipulating objects, and vocalizing commands and responses. In many ways, working with electronically augmented personal information infrastructures provides us with mechanical advantage, giving faster access to larger amounts of information, but the ergonomic consequences of such tradeoffs are an active area of longitudinal research (e.g., Rohmert, 1987).

Cognitively, electronic augmentations of personal information infrastructures allow us to access, manage, and communicate more information and more varied types of information. We are able to handle more complexity and process information more rapidly. These changes have also led to requirements for new skills in organizing information and for using technology itself, and consequently to additional cognitive pressures and stress. The interactions between cognition and technology are a central concern in this book.

Emotionally, technology has relieved some of the stress of communication in emergencies, but has likewise caused new pressures for improving intellectual productivity. On the one hand, computers have broken down interpersonal barriers of race, gender, age, and culture; provided new modes of expression; and opened up new levels of communication. On the other hand, computers have given bullies new avenues for intimidation and control; exposed many technophobics; and provided yet another excuse for excessive-compulsive perfectionists to avoid finishing their work. Although technological augmentations of personal information infrastructures provide obvious empowerments to disabled individuals, they have also isolated individuals who strongly depend on interpersonal interactions for satisfaction and joy. Powerful new environments must interact with human emotional evolution and very little attention has yet been given to the nature and consequences of these interactions.

PEOPLE AND ELECTRONIC ENVIRONMENTS: INTERACTIVITY

Human existence is a series of interactions with the environment. Most of humankind's early history was devoted to surviving these interactions and our science and technology is devoted to controlling the environment so that we might choose the interactions that serve personal or social purposes. We master and value interactions with other people through a variety of natural communication mechanisms that have evolved over time. *Interactivity*--the propensity to act in unison with external objects or other people--is a basic human characteristic. The complexity of modern society forces us to interact with more institutions and systems, using limited and cumbersome communications mechanisms we generally characterize by phrases such as "bureaucratic protocol." Interactions with institutions are becoming less natural because electronic systems are slowly replacing human "front ends" to the overall institution. Thus, our interactions with the environment are constantly evolving.

Information seeking is fundamentally an interactive process. It depends on initiatives on the part of the information seeker, feedback from the information environment, and decisions for subsequent initiatives based on this feedback. Our personal information infrastructures serve to regulate and standardize our interactions with information. The interactive nature of information seeking is described in detail in subsequent chapters, but this fundamental characteristic is necessary to understanding why electronic environments are so conducive to information seeking.

Interactivity has become a central characteristic of computers. The essence of computer programming is to apply computation and memory to control physical devices according to inputs provided by the user at "run time." The speed of computers allows programs to compare inputs to stored responses and to execute those responses almost instantaneously, thus giving the illusion of interaction. Interactivity is what allowed the personal computer market to develop and is the key reason that computers are increasingly important tools for group work and decision making. The electronic spreadsheet is a success not because it serves as an editor for numbers, but because it enables the interactive process of "what if" analysis (Kay, 1984). Human activities that are inherently interactive in nature will be strongly affected by computer technology, and information seeking is one such process. In addition to design principles such as giving the user control of a system through interface options, maximal interactivity in interface design is advocated. As a simple illustration, consider a system such as Perseus (Crane, 1992) that provides access to multiple images of objects such as vases. When a user has identified a vase of interest, a descriptive window is displayed that includes a list of all the views available for that vase. Rather than inviting the user to select a number of views (e.g., 1,2,7,34) to see in advance, maximal interactivity suggests that the user selects a view and then goes easily and quickly back to the vase description to select another view. This keeps the user engaged in concurrent viewing and decision making rather than batching the decision making first and then the viewing. Of course, the option to make multiple selections could also be provided, but the default should be to invite interactivity.

Psychological and sociological investigations of interactivity have been spurred by the development of computers and have helped to define the field of human-computer interaction (HCI). HCI aims to discover theories that explain interactions among humans and computers and to develop interfaces that support these interactions. One view of HCI is based on communication models. According to this view, the human and computer are senders and receivers of messages relayed along communication channels. The communication channel is called the interface and the goal of design is to develop high-bandwidth channels that reliably facilitate the flow of messages. This view is a useful one, although it is necessary to augment it by considering the context of the overall communication system. From this point of view, interface design is based on understanding fundamental features of humans and systems. Important features include: physical constraints such as memory capabilities, information transfer limits, computational ability (logic more than arithmetic); and conceptual constraints such as knowledge of the task domain and mental models for objects and processes in the world (especially mental models for the other receiver/transmitter and the communication process itself).

Using this HCI perspective, it is natural to study human-computer interactions by first considering human-human and computer-computer interactions. An implicit bias in most HCI research is that human-computer interaction should be based on models of human-human interaction. Human-human interactions are multimodal and complex, and there are rich traditions of evidence from the fields of communications, psychology, sociology, and physiology. However, human-human interactions are also ambiguous, are situated in that they depend on awareness of context, and depend on redundancy to be effective[2]. Computer-computer interactions are well-defined and fully understood from a technical perspective, but the many layers of translation needed to represent ideas as binary signals lead to the ultimate example of "losing something in the translation." As a result, computer-computer communication is best for explicit, discrete information transfer rather than complex or creative expression. A crucial design decision for all HCI is determining the degree of mix between human-human and computer-computer models. Consider some of the following differences between humans and computers with respect to information processing features.

Human memory is believed to be made up of working memory and long term memory (e.g., see Wickens, 1987). The size of working memory is severely limited (5-9 units or memory--chunks) and long term memory is infinite in size. Human memory is thought to be associational and episodic, and allows us to make direct connections between memory traces. Human memory is unreliable and we are not always able to recall information on demand. Computers have random access and mass storage memories as analogs of human working and long term memories. Computer memories are highly reliable (as long as physical requirements like electrical power or absence of electromagnetic disturbances are maintained), and their size has been increasing steadily. Computer memories are not associational but depend on pointers and indexes that specify exact addresses for information units. Also, all computer memory is stored in binary code, and thus far we have been unable to represent the full richness of events, let alone ideas. This severely limits the range of practical expression, especially with respect to nuance and impression.

Human computational ability has been estimated as 10/cycles per second[3], where each cycle is able to recognize or select some chunk of information. Delimiting the exact scope of what determines a chunk is an open problem in cognitive psychology. Computers have excellent computational power, capable of millions of computational cycles per second. Computer cycles are able to execute arithmetical steps or make comparisons between items in memory. Humans can execute a wider range of processes on a wider range of information units than computers, while computers can execute more well-defined processes with much narrower information units at much faster rates of speed.

Humans are able to receive and transmit information using a variety of channels. Output channels include: voice, gesture, facial expressions, and a host of muscle movements that control devices such as pencils, keyboards, musical instruments, and pointing devices. Input channels include: sight, hearing, touch, smell, and taste. All of these channels have associated bandwidths to meet environmental and intentional conditions. Humans are able to integrate these channels in parallel, although each channel may have strict sequential limitations. Bandwidth varies greatly[4] and the rich array of channels humans use is referred to as high fidelity. Computers are able to transmit at enormous rates of speed (e.g., Sumner (1990, p.14) predicted that in a few years it will be possible to send all recorded knowledge past your house in a second). However, computers currently offer low fidelity communications overall. A goal of

HCI research is to develop input/output devices and mechanisms that map more naturally onto human channels.

Humans develop highly integrated knowledge about areas of interest. Knowledge can be based on experience, acquired vicariously and symbolically, or created internally through computation, inference and imagination. Human knowledge is dynamic. It grows according to conscious activity and it ebbs according to a variety of physical and psychological states. Computers have highly rigid knowledge bases provided by the external world. Although computer knowledge bases are potentially massive, it is an open problem as to how well they will perform in application tasks.

Humans are learning systems. We apply our cognitive power of reflection to our continuous life experience and develop understandings of the environment--we think about what is presently occurring and what has happened to us and learn to improve performance in the future. We develop mental models for objects, events, and activities so as to benefit from subsequent interactions. Users of computers develop mental models that allow interaction, although these mental models vary in accuracy and depth. Although models for machine learning have been proposed, machine learning remains very much a long-term goal (Carbonell, 1992). Computers are given knowledge about the world by programmers. This knowledge includes primitive models for users that facilitate the human-computer interaction process. User modeling research aims to develop characterizations of human task performance and to build these characterizations into the design of systems, but at present these models are quite crude and coarse (e.g., Allen, 1990, Daniels, 1986).

Clearly, with such differences, communication between humans and computers will be more difficult than either human-human or computer-computer communication alone. A primary goal of HCI is to develop new devices and interaction styles that take advantage of the respective strengths of humans and computers so as to build new mechanisms and languages for human-computer interaction and collaboration.

Communication is perhaps a better metaphor for human-computer interaction than a precise model and a variety of alternative viewpoints are possible. Some people believe that communication is strictly a human to human activity and object to the anthropomorphic implications of human-computer communication. One possible strongly human-centered design consequence is to consider computers as prosthetics for humans. Interaction is limited to control and manipulation by the human, although preprogrammed actions can be initiated by specific sets of conditions dictated in advance. Quite a different viewpoint assumes that computers are a new type of entity with intelligence and evolving volition. One possible communication viewpoint is to assume that humans and computers should have equal authority for initiation and action. Still another viewpoint is task-oriented in that humans and computers are viewed as components in a larger system. Interaction may be viewed as collaboration between humans and computers where the task determines how best to take advantage of respective strengths and minimize respective weaknesses.

Two issues that emerge from consideration of models and metaphors for interactivity between humans and computers are related to autonomy and adaptability. Consider, for example, the degree of autonomy that should be built into robots and intelligent agents. Should a robot under the sea or in space be controlled through telepresence or should it be capable of making autonomous decisions based on environmental conditions? Setting aside the technical problems of both capabilities, and acknowledging that some combination of each is desirable, there are fundamental philosophical differences between arguing for autonomous versus directly controlled robots. The reality is that robots will be designed and built, and decisions about levels of control will be made on the basis of state of technology, cost, and task analysis. Similarly, agents that are proposed to assist humans in exploring information spaces will be designed from both autonomous and direct control perspectives.

It is naive to believe that any single interface will serve the needs of all users for all tasks. We go to people with special skills and knowledge to help us solve problems (e.g., physicians, attorneys, librarians) and we go to physical places to take advantage of special features (e.g., spectacular scenery, healthy climate, commercial activity). Interfaces should also vary according to information-seeking task and personal characteristics. Some argue that an interface should adapt to the user automatically based on user profiles or records of past experience; others argue that all adaptations must proceed from the conscious actions of users. Just as with autonomy, systems will be built that reflect both viewpoints. Rather than interfaces that "act human", we should develop interfaces that improve performance by reinforcing the characteristics that bring us to the system to begin with--we need more reality, not more virtual reality! For example, speech interfaces could be much more commonly used today if the quest for human-like continuous speech had not drained so much talent and resources. Short, command-like speech can be quite effective for many common tasks and can be easily provided in most interfaces as an inexpensive option.

HCI research has significant implications for information seeking because information seeking is such a critical process in an information society and because our personal information infrastructures are becoming increasingly dependent on computer technology. Although many of the issues of general interest to HCI research also apply to interfaces that support information seeking, there are particular requirements and conditions that must be considered. The objects of information seeking are ideas and their many representations. These abstractions are distinct from manipulating physical objects and typically are less well-defined than manipulating numeric or factual data. Interfaces will most likely need to be more personalized and flexible since information seeking depends so heavily on interactions among complex information-seeking factors.

Interactivity that is dependent on such interfaces raises new variations on old problems. Just as information sources vary in validity and reliability, so will interfaces, and users must develop evaluation and selection standards for interfaces just as they have for information sources. Whether these interfaces ameliorate or exacerbate the problems of information pollution (overload) remains to be seen. There is a tension between redundancy and memory--we need redundancy and context to remember information. When does context become information pollution? If our filtering agents become so efficient that we eliminate most redundancy and context, will we "know" anything?[5] What levels of adaptability are appropriate and "comfortable" for humans is also an issue of concern for researchers and designers.

HCI research is critical for meeting the challenges of information seeking in electronic environments. Devices and interaction styles must be developed that match the physical, conceptual, and emotional activities of accessing, assessing, and extracting information from electronic sources. To develop such interfaces, models of information seeking must be taken into consideration as the basis for design.

Of course, before these long-term interfaces are built, a host of specific design problems must be solved. Screen display layouts; interaction styles; mapping information-seeking tasks to levels of representation, mechanisms for controlling those representations; and mappings of physical devices to tasks are examples of immediate problems that are considered within the larger context of interfaces to support information seeking. A general problem with today's interfaces that support information seeking is that support is strong for some of the subprocesses but weak or non-existent for others. In the next chapter, the subprocesses will be considered in detail and interface issues discussed. The main problem is that today's interfaces focus on query formulation and results examination functions but ignore problem identification/clarification and information extraction (Marchionini, 1992). Much of what needs to be done in information-seeking interface design relates to perceptual versus cognitive processes. This book argues for systems that support active browsing and that minimize memory-intensive activities. Such systems amplify perception by maximizing interactivity, and augment cognition by freeing cognitive resources to focus on filtering/judging/interpreting information rather than attending to query formulation and system manipulation.

DEVELOPMENTS IN ELECTRONIC ENVIRONMENTS: SYSTEMS, DATA STRUCTURES, AND ALGORITHMS

Interactions with electronic devices have increased in all aspects of knowledge work, including information seeking. Developments in hardware, data structuring, and algorithms had early influences on information seeking by forcing experts to formalize information-seeking strategies. The computer systems of the sixties and seventies led mainly to analytical strategies which were based on careful explication of steps taken in manual environments and that took advantage of the power of electronic computation and storage. As will be argued in subsequent chapters, electronic environments of the eighties and nineties ushered in a new wave of expansion in electronic information seeking by allowing broader classes of information seekers to use highly interactive browsing strategies. Thus, the continued evolution of electronic environments has allowed the full range of information-seeking strategies used in manual environments to be applied and augmented. Today's storage, computation, and communication technologies allow full text and multimedia databases to be rapidly accessible by masses of end users in a variety of physical locations. The results of all these efforts are coming together to support the highly interactive information seeking described in this book. A brief overview of the developments most specific to information seeking follows.

Hardware developments have been dramatic and will likely continue as efforts in the U.S. such as the High Performance Computing Initiative (U.S. Office of Science and Technology Policy, 1991), and the many projects related to developing the digital libraries and the National Information Infrastructure; Europe's ESPRIT Program (Smeaton, 1992), and Japan's Fifth Generation Computing Initiative (Feigenbaum & McCorduck, 1983; Fuchi et al, 1993; see also the large-scale Japanese efforts to develop interfaces known as the FRIENDS21 Project, Nonogaki & Ueda, 1991) evolve. Computing power has grown from thousands of floating point operations per second (kiloFLOPS) to megaFLOPS to gigaFLOPS in the last decade, and parallel architectures are beginning to increase actual throughput for complex processing[6]. Most significantly, these dramatic increases in high-end computing have been accompanied by general availability of personal computers and workstations that provide substantial computing power to offices, schools, and homes. In addition to central processing unit power, developments in magnetic and optical technology make it possible to distribute large libraries of text and growing collections of images, sounds, and multimedia documents. Companion advances in display technology and data compression techniques allow electronic digital displays to surpass analog video displays and rival photographic technology. Optical scanner technology has become fast, accurate, and cheap enough to facilitate cost-effective conversion of archival or esoteric paper-based documents to electronic form and video capture technology offers similar capabilities for video.

Research on interfaces for online systems has made these systems more easily available by improving both learnability and usability. Advances have been made in user-system dialogues, i.e., by viewing computer use as a communication process; in assuring the consistency and clarity of messages, prompts, and feedback; by studying and improving command and interaction languages from the user's perspective; by developments in natural language processing; by considering documentation and instruction from the initial stages of design rather than as afterthoughts; by the use of metaphors such as the desktop to aid learning; by the development and testing of graphical user interfaces (GUIs) and window-icon-mouse-pointing systems (WIMPS) that facilitate more direct manipulation of data; and by the development of alternative rules and mechanisms for query specification and feedback.

The combined effects of hardware and interface research have yielded new genre of systems such as hypermedia that provide user-centered control of multimedia databases through application of powerful hardware and highly interactive software. These technical developments have not been made independently but represent the more general trend toward more efficient and effective automation of information work. Perhaps less dramatic, but equally important ultimately, is research on how information is collected, stored, and organized for eventual retrieval and use by information seekers. Research in document representation and retrieval techniques has been spurred by hardware and interface progress and is essential to the overall development of systems that support information seeking.

The traditional way to represent information documents in large collections such as libraries is to support search through indexing. Each document is assigned one or more index terms selected to represent the best meaning of the document. These index terms are then searched to locate documents related to queries expressed in words taken from the index language. In the simplest cases, the index language is a set of dates or names or identification numbers that serve as entry points to the database. This is the basis for database management systems where the documents are called records (tuples) and the index language is simply the set of possible values for key fields (attributes). The most important and difficult cases involve accessing information by "subject" and for this purpose, well-defined indexing languages that delineate concepts are developed. There are generic indexing languages for library access (e.g., Library of Congress Subject Headings), but specialized indexing vocabularies for specific literatures (e.g., Association for Computing Machinery Classification system for computer science, Medical Subject Headings for medicine, etc.) have been developed to index information in those areas better (see Soergel, 1985 for a full treatment of indexing).

The main technique used by today's large commercial information retrieval systems is to index each document by a number of terms (these range from a few to dozens) and create an "inverted file" for the database. The inverted file contains each word in the index language and pointers to each document indexed under that term. This technique usually is accompanied by algorithms that allow searchers to enter Boolean combinations of words as queries and to combine document hits appropriately and automatically. In the case of some online databases and most full-text CD-ROM databases, the index language consists of all words contained in the entire collection of documents[7]. Each document is then indexed by all words that occur in it. Full-text indexing is one way of supporting what is generally termed "string" search (since any string of characters can be located), one of the most significant differences between manual and electronic searching[8]. These techniques allow information seekers to find every document that contains any word they specify. The assumption underlying all these forms of indexing is that the "meanings" of documents and queries can be and are captured in specific words or phrases (see Frakes & Baeza-Yates,1992 for a collection of readings related to these and other information retrieval techniques).

Alternative approaches to representing documents have also emerged, including knowledge-based indexing languages that capture deeper and richer meanings of individual documents and the relations among documents. Humphrey (1989) has developed a frame-based indexing system for medical literature that augments the existing MESH and aides indexers in assigning valid and useful terms. In addition to improving document representation, such systems may also be used by the information seeker to augment terms specifically identified in a query with those implicitly related by the frame structures.

Another set of approaches to document representation treats the documents and queries as vectors (Salton, 1989; Salton & McGill, 1983; Salton & Buckley, 1990). Each cell in the vector corresponds to one term in the index language and in the simplest case, each value represents the degree to which that term occurs in the document. These values can be simple binaries (e.g., 1=yes, 0=no), raw number of occurrences, raw number of occurrences weighted according to the document length, or normalized according to frequency of occurrence in the entire collection. Queries are likewise represented as vectors and a variety of similarity measures can be used to match queries and documents (e.g., the cosine of the angle between the query vector and each document vector can be used as a semantic proxy metric to rank documents). The vector approach has significant advantage over traditional indexing methods for end users because retrieved sets of documents can be ranked, thus eliminating the "no hits" result so common in exact match systems. Experimental systems that provide ranked output have proven highly effective and commercial vendors have begun to offer ranked output features. Ranked output also provides a reasonable entry point for browsing.

Extensions of this statistical approach to information retrieval include models based on clustering techniques. Some clustering approaches compare vector representations in pairwise fashion to form groups of similar documents which subsequently are combined to form a smaller number of still larger groups. Alternative approaches begin with one or more key documents that focus on a topic or concept and then process the remaining documents by comparing similarity values and assigning them to the concept group most closely matched (see Rasmussen, (1992) for an overview of clustering techniques). A model known as latent semantic indexing has also been used to automatically process documents and queries. It assigns terms based on frequency but also uses correlational metrics to augment co-occurrences across documents. The technique uses singular-value decomposition, an algorithm related to factor analysis that collapses a large matrix into a smaller set of distinct factors (Dumais, 1988; Deerwester, et al, 1990).

Relevance ratings arise from judgments information seekers make about a document with respect to an actual problem. Since the query expressed in any search is a surrogate for the problem, models that match queries to documents assume accurate problem-to-query mappings for query to document comparisons. Probabilsitic models acknowledge that there are degrees of relevance and are based on some estimated probability of document-query relevance. These models rank documents according to system-estimated probabilities for relevance of queries and documents or query classes and document classes. User probability estimates also can be used as the basis for relevance feedback to the system (see Bookstein, 1985 for an overview of probabilistic retrieval models and Larson, 1992 for empirical comparisons of vector and probabilistic models).

In addition to matching based on vector similarities or probability rankings, connectionist models of pattern matching have been proposed. Models based on supervised neural network algorithms (Belew, 1986) or unsupervised algorithms (Lin, 1993) have been tested (see Doszkocs, Reggia, & Lin, 1990 for a review of connectionist techniques for information retrieval). These approaches take vectors as inputs and provide clustered sets of outputs. The main idea is to process the document collection by comparing an input vector (e.g., a randomly selected document vector to all document vectors to determine the best match. The vector weights for the "winning" document are then adjusted to more closely reflect the input vector, as are its proximate neighbors. This process, called "training," proceeds for thousands of iterations until the document vector space stabilizes (few weight adjustments are made between subsequent iterations). This space then represents documents in neighborhoods where proximity is dependent on document similarity.

All these techniques provide a basis for ranking documents which provides a good entry point for a user-centered, interactive model of human information seeking. Relevance feedback is a technique that has proven highly effective for improving retrieval (Harmon, 1992). Users examine results and indicate those that are most useful. The system then locates more documents that are similar and the process continues iteratively. Relevance feedback requires an interactive setting where information seekers select relevant items and the system reformulates the search. The reformulation may be based on simple matching of index terms used in and exact match Boolean system, or on new query vectors adjusted according to a variety of adjustment techniques. Together, ranking and relevance feedback support highly interactive information seeking. With continued developments in hardware and interfaces, these techniques offer the potential for new generations of highly interactive systems to support information seeking.

Most operational systems apply one combination of these approaches and users must learn the best strategies to use for that system. Some systems are more hybrid, offering information seekers choices between exact match Boolean and ranked approaches. As this trend continues, information seekers will need more guidance in selecting the best strategy to use for specific problems. As the next generation of analytical and browse strategies evolve to accommodate such systems, our personal information infrastructures will be augmented as well.

Chapter 2 Notes

1. The term "system" is used to include information objects such as books or people, as well as electronic objects.

2. Shannon and Weaver (1949, p.56) report 50% redundancy in ordinary English language, i.e., about half of what we write is determined by the structure of the language itself rather than the content we wish to present.

3. Card, Moran & Newell (1983) report cycle times for perceptual and cognitive cycles processors. The time to recognize distinct flashes of light ranges from 50 to 200 milliseconds, Card, Moran & Newell use 100ms as an average. Times for various object recognition, counting, and selection tasks vary from 25 to 170 ms and they use 70 ms as average. An estimate of 10 cycles per second is based on these data. Potter & Levy (1969) cite ranges from 50 ms to 300 ms for single pictures with low to high levels of visual noise, and their studies demonstrate that accuracy of visual recognition improves as display times increased from 125 ms (16% accuracy) to 1000 ms (80% accuracy).

4. Human speech (both speaking and listening) averages 120 words per minute (wpm) or roughly 80 bits per second (bps) (Streeter, 1988) . Human reading averages between 200 and 300 wpm, roughly 140-200bps (Streeter, 1988 cites 200-300wpm, and Hulme, 1984 cites 250-300wpm). If we consider the 100,000,000 rods and cones in the human eye as capable of accepting a bit in a "glance", visual input is approximately 100,000,000 bits per glance. It is important to note that these values have not changed over the course of the last 6000 years!

5. For example, much of our social intercourse depends on current events and "trivia" that we accumulate. More importantly, insights and research advances come as a result of humans making connections between seemingly unrelated ideas or events. The development of highly constrained filtering agents can have serious consequences for us individually as well as for civilization itself.

6. For example, the parallel architecture of the Connection machine has been applied to send a query concurrently to multiple databases in a wide area information server (WAIS), (Kahle & Medlar, 1991).

7. In actual practice, commonly occurring words (stop words) are not included and word stemming principles are often applied. See Belkin & Croft (1987) or Salton (1989) for an overview of information retrieval principles.

8. String search based on inverted files or on signature files (Faloutsos, 1985) is more properly called "full-text" or "word" search since in true string search each character or character group is scanned in linear fashion at the time of search. Scanning techniques are impractical in large databases since every character must be examined and compared; thus strategies such as indexes are applied in advance of search.