Chapter 5. Analytical Search Strategies

Necessity is the mother of invention

Complexity...

Storage and retrieval of scientific texts was one of the early applications of computers and by the early 1960's schemes for automatic indexing and abstracting had emerged (e.g., Doyle, 1965; Luhn 1957, 1958; O'Connor, 1964; Tasman, 1957). As online systems emerged in the 1960's and 1970's, more databases and new search features were created to give professional intermediaries more power in searching for information. Searching in online systems became complex and creative intermediaries developed systematic strategies for eliciting user needs, selecting terms, synonyms, and morphological variants appropriate to the user need and the system, using Boolean operators to formulate precise queries, restricting those queries to specific database fields, forming intermediate sets of results, manipulating those sets, and selecting appropriate display formats. The strategies and tactics professional intermediaries use are meant to maximize retrieval effectiveness while minimizing online costs. These strategies are goal-oriented and systematic and are termed analytical strategies. In this chapter, several analytical strategies are described to illustrate how electronic environments have changed information seeking by allowing searchers to systematically manipulate large sets of potentially relevant documents. These strategies in turn influenced subsequent design of online systems. Studies of novice users working with various online systems are described next, illustrating how difficult analytical strategies are to learn and apply and the need for electronic systems that support informal information-seeking strategies for end users.

ONLINE SEARCH STRATEGIES AND TACTICS.

Using an index to find target selections is the basic analytical search strategy. Such look ups depend on an ordered list of concepts that provide pointers to primary information. Textual indexes are often ordered alphabetically and depend on pointers such as page numbers, file offset, and record number. The strategy is to start with index entry points and follow the pointers until information is found or all entry points are exhausted. Electronic systems that are based on inverted files depend on this basic strategy. When Boolean-based query languages are added, users can create complex queries that allow multiple index entry points to be combined in one query. Likewise, users can create multiple sets of documents related to specific entry points and then combine those document sets to obtain the union, intersection, of difference of the sets. Various specific strategies have been developed for use in online systems.

The most widely used online searching strategy is the "building blocks" approach (Harter, 1986). During problem definition, the information seeker identifies the main facets or concept groups associated with the problem. These facets then become the basis for specific query formulations that retrieve sets of document citations for each facet. After the individual sets are formed, they are systematically combined with Boolean operators (most commonly AND) to produce a document set relevant to the problem as a whole. Thus, the individual facet sets (blocks) are combined to build a solution set for the problem. For the building block strategy, various tactical options may be used to build queries for facets (e.g., using controlled vocabulary, ORing synonyms, limiting to specific fields) and to combine resulting sets (e.g., combine sets in pairwise fashion or all at once).[4] Which tactics to use will depend on the information seeker's experience in using feedback from the system. For example, if a set for a facet is too small, the searcher may decide to combine it with another set or reformulate a broader query to increase its size. When the combined document set is obtained, inferences are then made about precision and recall and adjustments made to individual facet sets or to how the facet sets were combined. Figure 5.1 presents an example search that uses the building block strategy. The original statement of the problem was used as the basis for a high recall search (the recall search formed 68 sets and was substantially longer than the high-precision search, see Marchionini & Barlow, 1994 for full searches) and then the precision search was conducted based on additional conversations with the end user. Note that a specific syntax is required and that some additional tactics such as sifting intermediate result sets are used within the overall building block strategy that identifies document sets for the two main concepts and then combines those sets with the AND operator.

[Place Figure 5.1 about here]

The building blocks strategy is popular because it simplifies the process by breaking it down into manageable parts. This is so for both the conceptual analysis of the problem and in the technical specification and refinement of Boolean queries.

Another strategy that finds common use is the "successive fractions" approach (Meadow & Cochrane, 1981). This approach begins with a large subset of the entire database and successively pares it down with concepts specific to the problem. As with the building blocks approach, there are variations on the main strategy, e.g., Harter distinguishes three successive facet strategies according to how the initial set is obtained and how the successive facets are ordered. Hawkins and Wager (1982) note that the successive fractions strategy works well with problems that are vague or broad, and cite its advantages for backtracking and monitoring search progress. As with the building blocks strategy, successive fractions is popular because it simplifies the search process by breaking it into a sequence of systematic and discrete steps.

A third general strategy is the "pearl growing" approach (Markey & Cochrane, 1981). This method begins with a specific document or document set that is known to be relevant (a pearl) and uses the characteristics of that document to successively grow a set of related (and presumably similarly relevant) documents. Given an actual document or citation, the information seeker uses assigned index terms, title or text words, names, citations, publication data, or structural and statistical properties to formulate queries to retrieve subsequent sets. After locating the "pearl" to use, the most difficult problem with this strategy is determining when to stop forming and "growing" subsequent sets--it requires more active engagement on the part of the information seeker. Hawkins and Wagers (1982) point out that this strategy is highly dependent on interaction between the searcher and the system and note that for the case of using indexing terms from the retrieved pearl, is impossible to do in manual environments since indexing terms are seldom provided in printed indexes. Like the other strategies, pearl growing is dependent on searcher inferences and has many variations and extensions. Of particular interest in highly interactive environments are relevance feedback techniques that are based on the same underlying assumptions about relevant documents having common structural and semantic characteristics. Although this strategy begins with a well-defined and solid entry, it is less algorithmic than building blocks or successive fractions, and requires more searcher interaction with the system.

A fourth general strategy was characterized by Hawkins and Wager (1982) as 'interactive scanning". This approach requires high levels of user-system interaction and is less algorithmic and more like guided discovery learning. The idea is to begin with a comprehensive set of documents generally related to the problem area (e.g., retrieve a large set of documents using one or a few general terms). By scanning the documents, key features of the problem are noted (e.g., authors, terminology, methods) and these features are used to formulate and pose successive queries that further clarify the problem. As understanding of the problem progresses, documents or sets are printed or saved as part of the final resultant set. Clearly, this strategy requires continuous cognitive attention, changing criteria for judging relevance as the problem and its associated literature is better understood, and plausible reasoning about when to terminate search. Hawkins and Wagers note that the strategy takes more time than others but is useful for professionals conducting searches for high recall or for problems in domains unfamiliar to the end user or the searcher. It may also be used by professionals to explore a topic area before applying one of the more analytical strategies. This strategy is used by novices in many different environments and is much closer to browsing strategies then other analytical strategies.

Other strategies have been described in the literature of online bibliographic searching. Variations of a simple strategy, known somewhat disparagingly by professional intermediaries as "quick-and-dirty" or "easy" searches are often used by novice searchers and sometimes by professionals for background or intermediate information ("briefsearch" in Harter, 1986). Wagers (1989) describes the steps in the easy search as follows: select a database, write statement and divide into concepts, compose a simple query using a few terms linked by AND or OR, display an abbreviated record, modify with simple changes as needed, print complete records retrieved. He compared easy search strategies to sophisticated strategies (e.g., use field limitation, more synonyms, special codes) for the same search questions and found that average precision was similar across all searches, but sophisticated searches yielded higher recall on average. Not unexpectedly, the best easy search results were for problems with few concepts that were expressible with few words or phrases that were common in the literature. Modifications to the original query yielded large improvements and illustrate the importance of iterative and interactive searching regardless of type of strategy used. These results parallel those of Salton and his colleagues (Harmon, 1992; Salton, Fox, & Voorhees, 1983) who found that relevance feedback significantly improves retrieval outcomes. Overall, results of the easy search strategy were judged to be acceptable by end users.

Vigil (1983) described a strategy he called "closed-loop relevance clustering," which uses the NOT operator to successively remove redundant documents from sets formed as a result of query modifications and combinations. Once a reasonable set is retrieved, a second set is formed and NOT is used to find the difference between the two sets. If the difference contains no (or few) relevant records, search terminates, otherwise, a third set is formed and NOT is used to determine if there are relevant records in the difference between the newest set and the overlap between the original set and difference of sets one and two. The process continues until no new relevant records are found in the successive differences. Although preliminary results of tests demonstrated that the strategy yielded reasonable retrieval, its main contribution was to remind intermediaries of the uses and limitations of the seldom-used NOT operator.

These strategies are often combined in actual practice. For example, interactive scanning or citation pearl growing may be used to generate terminology for facets that eventually are used in a building block strategy. These strategies are influenced by the characteristics of online systems themselves. Boolean-based retrieval based on inverted file organizations, controlled vocabularies for bibliographic records, and charges based on connect time all deeply influenced the development of these strategies. For example, these characteristics lead to careful planning before going online. The strength of this influence is evident from consideration of how CD-ROMs have affected information seeking--end users are able to use informal strategies and librarians who teach users how to search have difficulty convincing learners that planning in advance is helpful.

Online search tactics.
Research and practice in online searching has identified the tactical actions that experts use in conducting searches. Although many of the tactics are usable in manual environments, many are made manageable or practical by electronic systems. Bates provided an early taxonomy of search tactics (Bates, 1979a, 1979b). She identified 29 search tactics in four categories plus 17 idea tactics. The search tactics include monitoring, file structure, search formulation, and term categories. They range in specificity from very general, strategy-like suggestions (e.g., break complex queries

down into subproblems), to more specific guidelines (e.g., use alternative affixes of a word), and include some common-sense, but often overlooked suggestions (e.g., see if the search has already been done by someone else). Her idea tactics include a series of psychological suggestions to initiate and further difficult searches. They range from general tactics like brainstorming or consulting with a colleague to specific suggestions like focusing on a broader or narrower formulation of the problem or stopping to do something else for a while. Bates' tactics have served to highlight the psychological aspects of online searching and have influenced both the training of professional intermediaries and system design.

Harter and Rogers-Peters (1985) presented a more detailed taxonomy of tactics for online searching. They identified 101 tactics in six categories: philosophical attitudes and overall approach, language of problem description, record and file structure, concept formulation and reformulation, recall and precision, and cost/efficiency. Attitudes and approaches stress general tactics such as willingness to browse and learn and specific suggestions such as knowing one's terminal well. The 26 language problem tactics are divided into five subcategories (general development of search terms, acronyms and abbreviations, spelling and usage variations, compound terms, and codes) and include suggestions to use thesauri, use index terms in retrieved records in reformulations, begin with most specific or relevant concept first, use antonyms, and check spelling and morphological variants. These tactics closely parallel Bates' term category tactics. Tactics in the record and file structure category include the general recommendation to always question null sets, and suggestions to know about field and record structures, stop word lists, default search fields, and parsing rules. Eighteen concept tactics in two sub-categories (formulation and reformulation) include cautions about using the NOT operator carefully, using Venn diagrams, ordering facets from most to least relevant, saving queries and strategies for future searches, and browsing retrieved records to aid in reformulation. Twenty tactics in four subcategories (specificity of concept definition within facets and of whole facets, narrowing, and broadening) are presented for increasing or decreasing recall and precision. Many of these are specific statements for Bates' general search formulation and term tactics. For example, tactics for narrowing a search include: limit search to specific fields; decrease use of truncation; use a classification code to limit to a subject area; limit search by date, language, or publication; and tighten word proximity. Finally, twelve tactics to decrease cost or improve efficiency are provided, reinforcing the overall importance of time and cost pressures for professional intermediaries.

Harter and Rogers-Peters' taxonomy is of immediate practical value to professional searchers because it is more specific to Boolean-based, bibliographic systems than Bates' more generic compilation. Both sets of tactics illustrate how electronic systems have influenced both the theory and practice of information seeking since so many of them are specifically enabled by electronic technologies. As with strategies, however, there are questions about how these tactics apply to full-text, interactive systems designed for end users. To what extent are the strategies and tactics artifacts of early, bibliographic database technologies? Which technological factors (e.g., retrieval techniques, interfaces, hardware limitations) and content factors (e.g., secondary or primary content, size, scope and organizational structure) most affect learnability and usability? Are end-users willing and able to learn and apply these approaches and techniques? Of the techniques most appropriate for end-user-oriented systems, which should be embedded in interfaces automatically so that users need not learn them and which are essential for users to master?

These results emerged in bibliographic online systems. As full-text databases become more widespread, new strategies and tactics are needed. Although commercial legal and other full-text databases support proximity limits and full-text queries, the economics of system development, the underlying data structures and retrieval algorithms remain the same, limiting the development of strategies more appropriate for full-text online databases.

Searching full-text online databases.

As full text and other primary databases become more widely available, expert information seekers adapt their patterns strategies, and tactics. Tenopir and her colleagues have been foremost in studying full-text database searching by experts (Tenopir, 1984; Tenopir & Ro, 1990). She and her co-workers found that full-text searches yielded higher recall results than searches limited to bibliographic records that included abstracts. Full-text searches were more costly, however, and yielded somewhat lower precision. She recommended using proximity rather than Boolean connectives to combine concepts and to combine both free-text and controlled vocabulary to gain good results in full-text systems. Thus, the tactics used in full-text databases as well as the outcomes appear to be distinct from those in bibliographic databases. Full-text systems may require experts to adjust their strategies and tactics and perhaps develop new ones, although more experience and study of these effects are needed.

Full-text systems may change the expectations of information seekers. Tenopir reported that many information seekers noted that a significant difference between bibliographic and full-text searching was that full-text searching allowed immediate assessments of relevance to be made during search. Full-text searching removes the step of locating actual documents, improving the probability that relevance judgments are accurate and stable. This characteristic should be a clear advantage for end users who can immediately make informed judgments and extract relevant information as they make those judgments. The balance between recall and precision may also change in full-text environments. For example, twenty pages (screens) of bibliographic citations may be far too much for users to examine, but twenty pages (screens) of primary materials may be far too little. These early results with expert intermediaries using full-text databases indicate that there are unique characteristics of primary search systems that will require new strategies and tactics for experts and novices alike.

Lessons learned from professional online searching.

Much has been learned about information seeking in general from studies of professional users of online bibliographic systems. This experience base has illustrated how electronic systems differ from manual systems, highlighted some of the special capabilities and limitations of electronic environments, and delineated specific strategies and tactics used in information seeking.

Katz (1987) described four main advantages of online versus manual searching: speed, convenience, depth of searching, and currency. First, searching an online database is clearly faster than searching a manual index, especially when the search spans several years and requires handling multiple index volumes. Second, searching online is more convenient since a variety of databases can be accessed from different sites, including offices and homes--the indexes come to the user rather than the information seeker going to them. Furthermore, traces and results can be easily printed and saved with online systems. Third, the quality of searches can be better in online environments since a variety of entry points can be used, possibly including text words in abstracts. More importantly, greater subject control is possible because terms can be combined with logical connectives. Also, searches can be automatically limited according to characteristics like date, type of publication or language. Although many of these information-seeking tactics can be executed in manual systems, it is highly unlikely that they will be applied as extensively. Finally, electronic databases are typically updated more frequently than printed indexes and thus more timely information can be obtained with electronic systems.

Online systems offer other advantages over manual systems. They can provide added value and adapt to the individual needs of different information seekers. In one form of added value, electronic indexes may contain more information than their printed counterparts. For example, the MEDLINE database, a bibliographic collection for the medical literature, is analogous to the printed Index Medicus. The online version of the database covers over 600 more journals than the printed version and about 65% of all the citations include abstracts, whereas none of the printed citations include abstracts (Grateful Med User's Guide, 1992). Electronic systems may also provide instruction and help for users. This assistance can be dynamic and context sensitive, and resemble specific advice. Online systems can also provide flexible or individualized interfaces for information seekers. For example, many systems allow users to directly specify commands or use menus to select options while searching. Information seekers with special needs can also be accommodated with alternative input (e.g., joystick or voice recognition) and output devices (e.g,, large print, graphics, speech synthesis). Thus, electronic systems provide windows of opportunity for designers to go beyond the limitations of bulky and static manual search systems.

Electronic systems are not without costs and limitations. Specialized equipment and power are required, users must be minimally computer literate, display technology is often inferior to paper for text and graphics, and costs of connection and access may be high. Nonetheless, these disadvantages are diminishing as computing pervades our work and economies of scale take effect. Regardless of their disadvantages, the advantages of online bibliographic systems over manual systems are rapidly extending to end-user-oriented, primary information sources such as full-text, electronic databases, statistical and graphical databases, and electronic forums and publications available via high-speed, world-wide communications networks.

The most significant contributions of research in professional online searching are related to our understanding of information seeking itself. Studies of online searching have led to the identification and specification of specific information-seeking patterns, strategies, and tactics, and illustrated the importance of underlying data structures, retrieval engines, interfaces, and user training for electronic systems. These studies have elucidated many interacting factors associated with information seeking and highlighted the complexity and richness of various steps in the information-seeking process. We have learned that expert intermediaries benefit from knowledge of specific controlled vocabularies and database structures; become highly skilled in specific query languages and the use of Boolean, proximity, and truncation operators; and are strongly influenced by training and connect charges.

In many ways, professional intermediaries have adapted their behaviors and expectations to the search systems made available to them. As new systems are developed and competition among different systems increases, new capabilities or relaxation of existing constraints may influence behavior and expectations. The strategies and tactics of expert intermediaries are commonly emulated in instruction for intermediaries and end users alike, and thus are perpetuated in today's more interactive and primary environments. It is important for researchers, designers, and users alike to ask: How do the strategies and tactics exhibited by expert users of online bibliographic systems apply to full text or other primary databases? Are they appropriate for end users? Which end users? Are they appropriate for hypermedia environments?

As systems evolve to include primary information and improved interfaces, more professionals have chosen to do their own searching and this has led vendors and agencies to develop specialized products and marketing strategies. This, in turn has led end users to develop expectations about the availability and use of information resources of all kinds. For example, medical students and law students typically take classes or workshops to learn about electronic search systems in their respective fields. These developments in end-user searching have begun to influence system design and the types of strategies and tactics information seekers use in their daily work. The early directions taken were to use training and to build interfaces that allow end users to apply analytical strategies. As more systems were developed and end user behaviors were studied, it became clear that the complexity of analytical search strategies often frustrated end users. As a result, systems began to emerge that are based on models of naive users and informal information-seeking strategies.

NAIVE MODELS OF INFORMATION SEEKING

Driven by growth in the personal computer industry in the 1970s and 80s, end-user computing became an active area of research and development. The information industry saw huge new markets for information products accessible by the increasingly large base of personal computers and thus explored ways to make databases more accessible to novice users. Likewise, the computer industry supported the development of interfaces that allow users with little specialized training to purchase and use computers. These forces allowed the human-computer interaction community to develop and promoted new innovations from designers and engineers. Studies of novice users of various types of information systems have begun to yield a more complete picture of information seeking in electronic environments.

Novice users of bibliographic search systems.

Because the end users of early bibliographic search systems were professionals in technical fields, there have been investigations of how these information seekers with high levels of domain expertise used these systems. Because medical information has life-critical consequences, medical knowledge is growing rapidly, and significant funding has been available, the field of medicine has figured prominently in these studies. Lancaster (1972), in an early study of the National Library of Medicine's MEDLARS search system reported that after training, medical practitioners were generally successful in locating information in the online system. Sewell and Teitelbaum (1986) reported similar results based on ten years of longitudinal data from various medical databases and search systems. The essential point of these results was that specific training was necessary to achieve appropriate results. To apply analytical strategies, users were required to learn at least a minimal set of search commands, simple Boolean logic, and some basic principles of database organization and medical subject headings.

An alternative to training end users to use complex systems is to develop specialized interfaces called "front-ends," that typically support only primary system features and offer users precise instructions and help (Meadow, 1992). Front ends typically aim to assist information seekers in using analytical strategies. A front end developed for professionals searching Department of Energy databases used iterative testing methods to develop and improve automated instruction and help (Meadow et al, 1989). Evaluations with end users of this system demonstrated that users conducted successful searches using the interface, although searches were overly complex and users had difficulty with term selection (Borgman, Case, & Meadow, 1989). Thus, customized interfaces designed for professionals who wish to access specialized databases have facilitated general success, although improvements are needed. These systems generally have taken the approach that judicious application of computational power (e.g., simplifying choices through menus or form fill-ins) will allow novices to apply expert strategies and tactics.

In another domain, broader access to the medical literature by physicians has been strongly encouraged by the National Library of Medicine through its development of the Grateful Med search software that runs on personal computer platforms and automatically connects to the NLM databases through phone lines or the Internet. First introduced in 1986, by late 1992 there were over 40,000 users generating more than 200,000 sessions per month with Grateful Med. Almost one-half of all searches of the Medline databases are conducted with Grateful Med rather than through direct connections using the MEDLARS language or other search front-ends. Although this has been a highly successful effort to make medical literature more accessible to end users, there are ongoing efforts to improve the system to overcome some of the problems users have with it. For example, almost one-third of all searches result in no documents. Incremental improvements in the Grateful Med system have been made over the years and a recent effort by a University of Maryland Human-Computer Interaction Laboratory team focused specifically on minimizing the number of "no hits" results and improving query reformulations. Most importantly, a case-based approach was used that provides different alternatives for continuing search depending on whether no hits or too many hits were found (Marchionini, Norman, & Boerner, 1992).

Another type of search system that has generated attention is the online public access catalog (OPAC). Online systems designed for domain specialists, OPACs must serve a diverse collection of novice and casual users. OPACs have allowed users more subject access, but many problems with OPACs have been described (Borgman, 1986; Bates, 1986). OPAC studies indicate that 30-50% of all subject searches result in no hits, many systems require users to use Library of Congress Subject Headings or other controlled vocabularies, and most systems offer no real browsing capabilities. Although aimed primarily at end users, OPACs have suffered from rigid underlying databases, primitive retrieval engines, and interfaces modeled on online bibliographic systems. This is beginning to change as OPACs are designed to circumvent these constraints, to accommodate specific user populations, or without the constraints from the beginning.

Rather than creating front ends for OPACs that lead users to learn and apply expert strategies and tactics, an alternative is to use a minimalist design approach (Carroll, 1990). This approach provides users with the bare essential features for basic functionality and progressively provides additional functionality as the user gains experience and confidence with the system. The Library of Congress took such an approach in designing its touch-panel-based interface for the combined LC catalog (Marchionini, Ashley, & Kortzendorfer, 1993). The ACCESS system assumed novice users who wanted quick access to the collection and had little expertise in using library classification systems or finding aids. The interface guidelines were developed in cooperation with the University of Maryland Human-Computer Interaction Laboratory and provided single-point access (e.g., a single subject) via a graphical interface and touch panel selections. In the Main Reading Room of the Library, ACCESS workstations as well as traditional workstations are available. The ACCESS workstations have been popular with patrons and with reference librarians who spend far less time helping patrons use the system than before the new interface was available. Users with expertise or with complex searches can use the sophisticated workstations which require considerable knowledge but allow complex analytical strategies that can save time. The system has proven so popular that other reading rooms of the Library have installed ACCESS workstations. This project is an example of an alternative front end that provides immediate usability but minimal functionality. Users who require more sophisticated access can move at their own comfort levels to the full-function workstations. Additionally, more functionality will be built into subsequent versions of ACCESS. Ultimately, the underlying databases must be changed to support more interactive end-user access, but providing distinct levels of interface for different types of users makes good sense in public areas serving a wide variety of information seekers.

A system developed for use at the Denver Public Library uses a highly interactive color graphics interface that provides multiple types of access to children. In addition to using a keyboard to enter subjects, users can select topics from menus for a specially constructed subject classification based on categories (e.g., animals, famous people, science, etc.) that children use and understand. Additionally, children can select categories such as new books, scary stories, etc. In addition to bibliographic information, users can see the book cover to help determine whether to go get the book from the shelf. This type of design is much more user-oriented than library-oriented and illustrates the user-oriented design paradigm.

A similar system was developed at UCLA (Borgman, et. al., 1990). This system uses an adapted Dewey Classification System to support subject searching. The system uses a bookshelf metaphor to support browsing and locational maps to assist children in finding the actual shelves in the library. Results of testing illustrate that all children were able to use a hierarchical browsing version of the system but younger children (age 7-11) had difficulty using a command-driven keyword version.

A radical departure in OPAC design is represented by the OKAPI system. This system uses relevance feedback and has virtually eliminated the "no hits" problem. Hancock-Beaulieu (1992) reported that the automatic query expansion feature of OKAPI has improved user success and that as many as one-fourth of the searches conducted by users would have failed in an exact-match Boolean system. As new OPACs are developed, designers are well-advised to build retrieval engines based on vector representations for documents so that ranked output and relevance feedback can be easily made part of the interface. Likewise, interfaces that are customized for particular user populations will also be more successful and satisfying for users.

Students and other end users without expertise in either specific domains or with retrieval systems have begun to use online bibliographic systems other than OPACs that assume analytical strategies will be used. Borgman (1986) studied how college students learned to use a search system for a subset of the OCLC bibliographic database by comparing conceptual and procedural training methods. Her research demonstrated the superiority of conceptual training for complex search tasks, but also noted how difficult learning and searching using Boolean queries was for novices. Huang (1992) studied college students who were trained to use an online search system to access various databases to find information for problems related to their own interests. Her research focused on subject pausing behavior. Presumably, a pause in activity reflects mental activity, including pondering how to use the system. She found that subjects were able to successfully find information related to their problems and that their pausing behaviors grew shorter as they became more expert with the system. Her work illustrates the learning curve of end users and provides another metric that may be useful for adaptable interfaces since systems can automatically gather pausing data as users work. Neuman (1993) investigated how high school students learned and used an online search system to access various databases to find information for school projects. Her interviews and observations demonstrated the difficulty students have in learning to use such systems and their frustration with not having easy access locally to the primary articles and reports that their searches located. All these studies illustrate how difficult it is for novices to learn and use existing online bibliographic systems and demonstrate the need for simpler and more effective interfaces and better access to the primary information found through bibliographic searching.

These studies demonstrate that given time and effort, novice users can become effective at locating pointers to relevant information by using online bibliographic systems and analytical strategies. This is especially true if users have expertise in the domain or are strongly motivated to persist. If novice users participate in training, they exhibit patterns of behavior, strategies, and tactics that strongly reflect that training. When training is absent or ineffective, novices exhibit passive patterns by accepting defaults and applying informal strategies and tactics. Improved interfaces or "front-ends" are somewhat effective in facilitating access and minimizing the technical details of query formulation, and more end-users are using online bibliographic systems to locate potentially pertinent information. More improvements are needed, especially in light of the improved interconnectivity brought by wide area networks.

Novice users of systems containing primary information.
Because studies of professional intermediaries using full-text systems have begun to demonstrate that different strategies are needed in secondary and primary search systems, it is reasonable to expect that similar differences occur for end users. A brief

overview of three technological developments is helpful to understand the results of the studies of end users that follow.

Technological basis for primary search systems. Although full-text, numeric, and other primary databases are available through traditional online services, there has been huge growth in primary search systems that are locally attached to end users' workstations. A *primary search system* is one that potentially provides direct answers to information seekers' questions or information that allows them to solve their problem. The most common example of a primary search system is a full-text database that provides firsthand rather than pointer (e.g., bibliographic) information. Other primary search systems provide statistical data, scientific datasets, images, or sound recordings. Growth has been spurred by developments in storage, display, computational power, and software advances that facilitate new types of electronic search systems. Increased computational power in the form of faster central processing units and expanded random access memories facilitated faster processing of large databases, supported highly interactive graphical user interfaces, and allowed information retrieval techniques used in large-scale environments, such as fully-inverted indexes and hypertext linking, to migrate to personal computer platforms. In addition to retrieval methods discussed in Chapter 2, three technologies are particularly important to primary search systems and warrant brief discussion.

Mass storage in both magnetic and optical forms allows large data sets to be accessed through inexpensive personal computers. Reference collections were obvious first choices and electronic versions were quickly produced for most basic collections such as dictionaries, encyclopedias, and popular directories. These were closely followed by textual collections (e.g., the Bible) and government data (e.g., the U.S. Census). CD-ROM databases have been particularly important in sparking development of full-text databases. Access to primary information on CD-ROM has proven popular with users and libraries are surely the largest market thus far for CD-ROM databases; in turn, this has affected libraries in fundamental ways. Subscription costs, workstation costs, and space requirements compete with traditional collection development and patron service costs for shares of library budgets and reference librarians are spending more time helping patrons use search systems. The most immediate impact on information seeking is that time pressures due to online connect charges are removed, thus allowing more interactive strategizing. Although CD-ROM technology makes full-text searchable material available to end users, it is not an ideal technology. Optical access speeds are much slower than access from magnetic disks and although the typical 600 megabyte capacity is large, many applications far exceed one disc, leading to awkward disc swapping procedures for users.

Bit-mapped, color displays gave new dimensions of representation to designers and led to graphical user interfaces (GUI) that facilitated easier use of these primary search systems. In a single decade, display technology on personal computers went from monochrome, 40 column by 20 row text-only displays to high resolution, color graphics displays that support multiple text fonts and styles and motion graphics. These developments have allowed designers to create direct manipulation environments for all applications, including information systems. For example, Young & Shneiderman (1993) created a filter-flow query system based on graphical representation of water flowing through pipes that allows users to visually manipulate data and see the intermediate results of Boolean queries. They demonstrated that this system was superior to a SQL interface to a database, thus making analytical strategies more manageable for end users. Display technology continues to advance, driven by portable computers and progress toward large, high-resolution flat-screens, however, these changes are costly to users who must upgrade or replace systems to take advantage of the new capabilities.

The confluence of work in text processing, human-computer interaction, and information retrieval led to the development of hypertext systems that allow users to move among various units of information (nodes) by following links. Hypermedia refers to hypertext systems that support multimedia nodes. See Barrett, (1988), Berk & Devlin (1991), Jonassen(1989), Nielsen (1990a) or Shneiderman and Kearsley (1989) for overviews of hypertext and hypermedia technology. Nodes can be small or large amounts of text, graphics, or any discrete object and links can emanate from anchor points in a node to anchor points in other nodes. Although most systems allow any node to have many in and out links, any specific link typically points to only one node. Selection mechanisms such as iconic buttons or embedded menus denote links and allow users to follow links by pressing keys or mouse buttons. Hypertext gives more control (and responsibility) to the user by allowing moves among nodes in non-linear fashion. This technology offer users new ways to use primary materials and has been applied to a variety of problems in documentation, reference, education, entertainment, and data management. Some problems associated with hypertext include disorientation for users and additional overhead for authors in planning and organizing their work. Hypermedia technology is often used by designers to provide users with alternative information-seeking strategies that are informal and interactive.

Studies of novices using primary search systems. Search systems with primary information were aimed at the end-user market and first appeared in schools, libraries, and other publicly-accessible sites. As part of ongoing investigation of the interactions of novice information seekers with primary electronic search systems, the author and his co-workers at the University of Maryland conducted studies with a variety of users and systems. Because they were instrumental in developing the information-seeking framework presented in this book, these studies are used here rather than a broader representative sample. A series of studies focused on fact-retrieval tasks for full-text encyclopedias and hypertexts. In one study (Marchionini 1989a), 28 third and fourth graders and 24 sixth graders conducted searches using a full-text, CD-ROM encyclopedia. After demonstration and practice sessions, students were assigned a fact retrieval question and an open-ended question and observed as they searched. Keystrokes were captured and together with observer notes, formed the basis for analyzing information-seeking processes. Results showed that older searchers were more successful in finding required information and took less time than the younger searchers. Although no differences in the total number of moves were found between the two groups, transaction matrices of the two groups showed that older subjects favored examination moves and younger subjects favored query formulation and refining moves. Some subjects posed phrases or sentences as queries, indicating that although they were forming mental models for the electronic encyclopedia that were distinct from the familiar print encyclopedias, they tended to overestimate the capability and "intelligence" of the electronic system. Most subjects accepted system defaults and limited themselves to Boolean ANDs as connectives. Analysis of search patterns showed that all these novices used heuristic, highly interactive search strategies rather than carefully planned analytical strategies.

A subsequent study (Marchionini, 1989b) focused on how high school students move from a print to an electronic encyclopedia. This investigation probed how 16 students conducted simulated "mental" searches, and searches in print and electronic versions of an encyclopedia over three sessions. Before starting each search, students were asked what they already knew about the problem and what they expected to find. Results suggest that encyclopedias are default search systems for many of the subjects, subjects knew and expected to find factual (what, when, where, who) rather than explanatory (how, why) information for various problems, students were able to generate vocabulary for queries beyond that found in the statement of the problem, and lower ability students had generally higher expectations for the electronic system. One-third of the subjects simply used the electronic system like a print system, entering an article via article title and reading the article in linear fashion. The other two-thirds of the students, however, took advantage of the full-text search features of the electronic version by adapting their print-based mental models to the new system.

Only half of the subjects in this study used AND connectives and none used OR, NOT, or proximity features. Subjects took almost twice as much time, posed more queries, and examined more articles in the electronic system than in the print system. The number of articles examined ranged from zero to six for the print searches and one to ten for the electronic searches, although differences in mean numbers of articles examined were not reliably significant across print and electronic searches. Equal proportions of print and electronic searches (20% and 19% respectively) yielded no hits and 10% of the electronic searches yielded too many hits--defined as students immediately reformulating a query before examining any articles. The proportion of no hits in this study is below the 30% range found in OPACs and other bibliographic system studies, reinforcing the distinctions between searches in primary and secondary databases. Although few of these students were observed to use the highlighted query terms in articles as browsing aids, some subjects commented on the usefulness of highlighted terms. Overall, these subjects performed somewhat perfunctory searches without taking full advantage of the interactive nature of the system. They were possibly influenced by the novelty effect of participating in an experiment and using what was then new technology. These results demonstrate that information seekers must be guided in adapting their mental models of manual systems if they are to take full advantage of the features of electronic versions. When introducing electronic analogs of manual systems, examples of how the new system is unlike the manual are as important as examples of similarities to promote learning.

Based on these earlier investigations, a study was undertaken to compare how novices used highly interactive browsing strategies and formal analytical strategies with an electronic encyclopedia system (Liebscher & Marchionini, 1988). Twenty-six ninth-grade science students were randomly assigned to either a browse or analytical treatment group and assigned the task of writing a short essay about the effect of the earth's rotation on climate and ecology. The browse group was trained to use a "scan and select" strategy: enter a simple query and scan the list of titles for potentially relevant articles, then use article outlines, headings, and highlighted query terms to quickly reject the article or extract relevant information. The analytic group was trained to use Boolean connectives to formulate precise queries to retrieve only a few article titles. Students conducted individual searches and all keystrokes were logged. Copies of the essays were subjected to content analysis by comparing numbers of prepositional phrases, numbers of relevant prepositional phrases, and grades assigned by a teacher.

Although no statistically reliable differences were found, predictable trends were apparent. Students in the "scan and select" group used fewer terms and retrieved more titles than those in the analytical group, but there were no differences in the number of relevant articles retrieved by the two groups (both groups had similar but poor precision). The analytical group showed greater within-group variance on these measures and reinforced the observations that the analytical strategy was more difficult to learn than the scan and select strategy. Essays produced by the analytical group were generally higher by almost one grade than those produced by the scan and select group. The scan and select group used more prepositional phrases per essay (26) than the analytical group (23), but fewer relevant prepositional phrases (18 and 23 respectively). A large negative correlation was found between essay grade and number of nonrelevant prepositional phrases. Students who used the scan and select strategy may have tried to incorporate more information since they scanned a great deal of text during their searches and were unwilling to give up large volumes of it. This may be related to the cognitive biases of representativeness and availability described by Tversky and Kahneman (1974). Another possibility is that the scan and select strategy may not have required students to organize facets of the problem during query formulation and thus led to difficulties in discriminating between relevant and nonrelevant information.

Overall, this study demonstrated that both types of strategy were effective in identifying relevant information but more attention must be given to finding ways to make "scan and select" strategies more discretionary. One approach may be to develop techniques to focus attention on how the problem relates to an entry point for browsing and then to continually relate the problem to text while browsing. The former technique is a type of query formulation step and the latter an examination step. Another approach is to provide training to users. Alternatively, diagnosis rules may be discovered that could allow the system to point out biases to users. The most essential result is that regardless of whether or not strategies are intuitive and easy to apply, information-seeking expertise includes knowledge of not only how to apply strategies but also their limitations.

These results with electronic encyclopedias led to subsequent investigations of novices using hypertext databases designed to invite browsing strategies (Marchionini, 1987). In an early set of investigations, paper and hypertext versions of a database and two hypertext access methods were compared (Wang, Liebscher, & Marchionini, 1988). The hypertext system (HyperTies) is distinguished by its intuitive interface and ease of use. In the first investigation 24 graduate students were randomly assigned to one of three versions of a database consisting of 106 articles about the Holocaust. One version was a paper version, one was a hypertext database with only an alphabetical index for the article titles and one was a full version of the hypertext with embedded menus (highlighted hypertext links) in the text as well as the index. All subjects conducted six searches. Because most participants were successful in finding relatively straightforward answers, there were no statistically reliable differences between the groups on accuracy of answers, mean number of articles viewed, and judgments of task difficulty. Subjects in the electronic groups judged the system to be slightly more difficult to use that the subjects using the paper version and they were generally less satisfied with the system, especially with respect to level of comfort and level of frustration. Subjects using the paper version were statistically reliably faster than the subjects in the electronic groups (means of 620 and 870 seconds respectively). All but two subjects in the full version treatment exclusively used the index rather than the hypertext links. Thus, users were not willing to adopt the novel hypertext jumps when the well-known index was available. This study illustrated some of the problems of using even a hypertext system optimized for ease of use when compared to familiar paper-based text.

In the second investigation, 36 graduate students were randomly assigned to one of two access method treatment groups. Both groups used the same database and full version of the hypertext system, but one group received training only in using the embedded menus and the other only in using the index. For the purposes of this study, use of embedded menus was considered a type of browsing strategy and use of the index as an analytical strategy. All subjects conducted six fact retrieval searches. Dependent measures included: success, number of moves made, total time to complete searches, number of articles viewed, and number of screens viewed. Subjects in the index group performed marginally better than those in the browse group on four of the five dependent measures (all but total time), however none of these differences were statistically reliable. When trends across the six questions were examined, differences in performance disappeared as more questions were completed. Thus, a learning effect was noted as subjects gained more experience with the browse strategy. These results may also have been biased by the subject population which consisted of library science students who had substantial experience building and using indexes. Subjects in the index group were generally more satisfied with the system, reliably so for ease of use, speed of use, and frustration level. Subjects reported their computer experience before participating in the study and were assigned to a low or high experience category. Only total time taken differed reliably when the two experience groups were compared on the five dependent measures. Thus, computer experience was not a predictor of performance. This study reinforced the earlier results that browsing strategies can be effective, but illustrated that even simple information-seeking strategies require some introduction and practice before they will be adopted.

A subsequent set of investigations was conducted to determine how access methods influenced information seeking and what design guidance could be discovered to guide adaptable interfaces. In one study, undergraduate subjects conducted fact-retrieval searches using a print based encyclopedia and three different electronic systems: a general electronic encyclopedia, a science and technology electronic encyclopedia, and a hypertext database (Marchionini & Liebscher, 1991). Subjects executed statistically-reliably faster searches in the print encyclopedia than in any of the electronic systems, illustrating the additional cognitive load needed to manage the systems that in spite of minimal training were still novel to them. Those in the electronic conditions examined three to four times as many articles as those in the print condition. Subjects in the hypertext treatment group outperformed those in all other groups in locating correct answers and executed about as many queries as those in the print condition. Those in the two electronic encyclopedias executed two to three times as many queries as those in the print and hypertext groups. This study illustrated that the hypertext environment required users to allocate less cognitive load to the system than did the Boolean-based full-text electronic encyclopedias.

Another study examined access method in detail. Twelve undergraduate students conducted searches in a hypertext over five two-hour sessions using different access methods (Liebscher, 1992). The database was an electronic version of a book on the topic of hypertext (Shneiderman & Kearsley, 1989) and five electronic versions were used: a version with only an alphabetical index, a version with only a subject index, a version that only allowed string search, a version that only allowed hypertextual links, and a version that allowed all four access methods. Subjects conducted individual searches for each of the methods for a total of 60 searches. They also conducted simulated searches verbally. Results showed overwhelming preference for string search as an access method. In spite of extensive usage, subjects had relatively poor mental models for the different access methods (e.g., thought that string search was actually subject search, thought the alphabetical index was less detailed than the conceptual index). These results illustrate that even if adaptable interfaces are provided, users may select and continue to use simple or default alternatives rather than those that optimize the task and conditions at hand.

Another environment that has proven fertile for the study of information seeking is the Perseus hypermedia corpus of materials on the ancient Greek world (Crane, 1992). This system is an aggregation of texts, images, and programs published as a set of HyperCard stacks and data files. Created for humanities scholars and students, the first release of Perseus consists of Greek texts and English translations for 10 authors; an historical overview of ancient Greek culture with explicit links to texts, maps, and images; approximately 7000 8-bit color images of vases, sculpture, sites, architecture and coins; textual descriptions for all objects; a Greek-English Lexicon and a morphological database for words in Greek texts; an atlas of Greece and the surrounding region that allows users to locate 800 sites; and an encyclopedia of art, archaeology and architecture. Tools for accessing and navigating these materials include: an index of all English definitions for Greek words; catalog and keyword indexes for all objects; string search for Greek and English words, a path tool that allows users to save, annotate, and edit tours through the database; and a set of menus that link various components of the database. The system represents a large and complex set of primary materials and tools for access and manipulation of information. During the development of the system, users were studied in a variety of college courses at different sites to determine how Perseus use influenced teaching, learning, and research[6]. These studies used observations, interviews, transaction logging, and document analyses to study how instructors and students used Perseus to teach and learn topics in Greek literature, ancient religion, archaeology, Greek language, and Greek culture and history.

Given its size, design, and the uses to which it was put, information seeking is a primary activity of Perseus users. For any given search task, there are multiple ways to locate a target and likely to be multiple targets as well. In our studies we were interested in how users learned to use the system, how the integration of texts, images, and hypertext tools influenced use, and how a complex hypermedia corpus affected teaching and learning (Marchionini & Crane, 1994). The main results of four years of evaluation follow.

Perseus offered mechanical advantage to instructors and students. It offered instructors an easily accessible and dynamic alternative to slides and transparencies for augmenting lectures. Furthermore, it offered a construction and delivery platform for instructors to provide integrated paths of texts and images that students could study individually or in small groups. From the learner's point of view, it speeded up word lookups for language translation, freeing the user to reallocate time to other tasks such as reviewing previous passages or looking ahead. Perseus also afforded learners exhaustive sets of passages containing specific Greek or English words, thus providing richer sets of evidence for supporting their interpretations and arguments. Students using Perseus were found to cite statistically reliably more passages and a wider variety of passages in their essays than students who did not use Perseus for their essays.

As with the previous studies of high school students using an electronic encyclopedia, flexible and interactive access to large volumes of material did not necessarily lead to better essays. Just as the high school students who used browsing strategies used many nonrelevant prepositional phrases in their essays, students who used Perseus rather than paper texts cited statistically reliably more passages in their essays but this did not lead to higher grades. Likewise, although students in a Greek language course could conduct word lookups somewhat faster using Perseus than a paper lexicon, this did not lead to superior translations. Although the mechanical advantages Perseus afforded students in finding text passages and translating Greek texts were not alone sufficient to produce superior translations or essays, Perseus use did allow some students to produce superior arguments.

Perseus use did enable new kinds of teaching and learning. Instructors and students often noted that the integration of text and graphics materials expanded the scope of course assignments and discussions, and students in literature courses reported using the graphic materials in Perseus to provide context for their textual interpretations. The graphical materials in Perseus were clearly a significant addition to teaching and learning and transaction logs showed that much of the time was spent locating and viewing images. The path tools in Perseus enabled new representations for student interpretations of the ancient Greek world, allowing them to present their ideas by weaving textual and graphical primary evidence into their paths. Students who knew no Greek were able to use the Greek-English lexicon to probe the meanings and contexts of key English words to discover how the ancient Greek authors used these words and thus what the concepts behind the words meant 2500 years ago. Thus, students without philological training were able to use the methods of the philologist to get closer to the ancient culture and to develop appreciations for the difficulties and biases of translations. Some students made interesting discoveries while exploring the database in open-ended fashion and these became the basis for their papers. Students also reported that Perseus encouraged them to use a broader range of materials in their studies.

The design and implementation of the Perseus system itself also affected student performance since it required concerted effort and time to learn if all its features were to be used[7]. Interface factors were more closely related to performance than were previous computer experience, illustrating the importance of well-designed interfaces for hypermedia systems. The design decision to use implicit rather than explicit links made Perseus somewhat more difficult for novices to use than some hypermedia systems that lay out explicit instructional paths.

These results parallel those in earlier studies of information seeking in electronic encyclopedias and hypertexts in that browsing and interactive search strategies can be effectively used, but users must apply them judiciously rather than mechanically.

Lessons learned from novice users of primary search systems.
Perhaps the most obvious result from studies of end users of primary search systems is the great diversity of abilities, characteristics, and experiences end users bring to these systems. A single information system may be used by children or adults, by users with varying amounts of time and patience, in public or private settings, and for a variety of information problems. Clearly, these systems must offer robust interfaces to serve such diverse ranges of users and uses. Nielsen (1989b) conducted a meta-analysis of 30 hypertext usability studies and found that individual differences and task were the two most important usability factors. These requirements have led designers to search for alternative or adaptable interfaces that may allow systems to serve broad user markets. Although user and task factors are most critical to information seeking, the interface is also important. For example, previous computer experience was not a strong predictor of performance but the learnability of the interface was found to be a performance-related factor in our Perseus studies.

It is evident from these studies that end users apply more naive information-seeking strategies and that interactive systems with primary information invite such strategies. These information seekers use browsing strategies and simple string search liberally in all types of electronic environments. Analytical strategies are more difficult to learn, although once learned, they typically yield more efficient results. Clearly, there is need for better understanding of browsing as an information-seeking strategy and development of interfaces that promote and support highly interactive strategies.

Information seeking in primary databases is more directly related to an end user's information problem and allows users to make relevance judgments on-the-fly as they interact primarily with potential answers to their questions rather than pointers to possible answers. This situation is more fluid than using secondary sources or a human intermediary. This is a boon for users but changes the meaning of relevance from a system-centered to a user-centered perspective. Relevance becomes totally dependent on sequences of examination and continual assessment by the user rather than on inferences about discrete sets of intermediate results. This makes it even more difficult to assess system performance via traditional recall or precision measures.

Results from studies of end users searching primary systems illustrate the importance of feedback. User expectations influence what strategies and tactics are used, and results in turn influence subsequent expectations. User expectations are often unreasonable, rooted in cognitive biases or misconceptions about technology in general. Children sometimes posed natural language questions to an electronic encyclopedia and adults had difficulty judging the size of a database. Users also tended to rationalize the usefulness of their results by including nonrelevant citations or clauses in essays. Feedback from the system must be carefully considered by designers from the start so that user expectations can quickly become congruent with system capabilities and resources. This is especially true in highly interactive environments where many cycles of feedback take place in single sessions.

Electronic systems have obviously affected the moves that expert and novice users make while information seeking. They have made tactics such as string search and manipulating intermediate sets of documents with Boolean operators much more practical. Analytical search strategies have grown out of expert users' taking advantage of Boolean operators and database structures. Experts can use them effectively, better interface techniques are making them more manageable, and as information seeking in electronic environments becomes more ubiquitous, better training will be provided to children. Experience with end users have repeatedly shown that analytical strategies are difficult to learn and use, and that informal, interactive strategies are preferred. Additionally, the emergence of primary search systems that offer interactive interfaces promotes such strategies. Electronic systems have begun to change user expectations about what is possible in an information society. Although it is too soon to say how electronic environments have affected general information seeking patterns, early indications are that the way we think about and react to information seeking is changing as more of these systems become available to us.

Chapter 5 Notes.

1. Combining sets in pairwise or other than all-at-once fashion is considered by Harter (1986) as a distinct strategy called pairwise facets.

2. A series of technical reports on the project evaluation are available as Perseus working papers #8, 9, 11, 12, and 14, 15, 16, and 17 (Perseus Project, Harvard University).

3. Perseus is somewhat unusual in that there are few explicit hypertext links but rather a series of indexes and tools that enable users to create their own links as needed. Thus, Perseus is much more like a library than a computer-based instruction system. This design philosophy is expressed by the system architects in Mylonas (1992) and Crane (1988).