User-Centered Evaluation of Information Retrieval

ABSTRACT

This paper briefly summarizes the history of evaluation in information retrieval and describes both the strengths and limitations of traditional criteria for retrieval effectiveness such as precision, recall, cost, novelty, and satisfaction. It presents a continuum of approaches to studying the user in information retrieval, and suggests that because the situations in which information is sought and used are social situations, objective measures such as retrieval sets and transaction log data may have limited usefulness in determining retrieval effectiveness. Information retrieval evaluation has been locked into a rationalistic, empirical framework which is no longer adequate.

A different framework of analysis, design, and evaluation that is contextual in nature is needed. User-centered criteria employing affective measures such as user satisfaction and situational information retrieval must be incorporated into evaluation and design of new information retrieval systems. Qualitative methods such as case studies, focus groups, or in-depth interviews can be combined with objective measures to produce more effective information retrieval research and evaluation.

INTRODUCTION

Linking Information Retrieval and Libraries

The key to the future of information systems and searching processes ...lies not in increased sophistication of technology, but in increased understanding of human involvement with information. (Saracevic & Kantor, 1988, p. 162)
Librarians are committed to assisting the user in obtaining access to the best materials available quickly, easily and efficiently, yet when librarians step aside from the reference encounter and let users pursue the information needed "on their own," many users fail utterly, or at least fail to achieve optimal results. Because of limited understanding of the information search process and even less understanding of how to evaluate that process, librarians may well wonder, "What is it that we are supposed to be helping the user to do?" and "How will we know when we have succeeded?" When the information search process involves machines, the picture becomes even more complicated.

In many libraries today, the intermediary role of the reference librarian is substantially reduced or nonexistent. One response to the invasion of end-user search systems such as online catalogs, database gateways, and CD-ROMs is to increase the commitment of effort and resources to bibliographic instruction (BI). This renewed interest in BI is reflected in conference themes, in the literature, in job descriptions, and in library school curricula. Unfortunately, much of the BI that is being done today is one-to-one or small-group instruction which is exceedingly labor-intensive and expensive. And despite the widespread interest in BI, there is very little evaluative data about its effectiveness.

Another response is to design systems that can substitute for the librarian as either an intermediary or as an instructor. This response represents a challenge of a different sort, one that requires enormous capital outlay at the outset, and goes well beyond the "help" screens that assist the user in attaining a minimal level of competency with system mechanics. These systems must not only perform adequately as systems, they must also "stand in" for reference librarians, assisting with question negotiation and clarification, and providing the friendly support and helpfulness that is associated with reference work. Unfortunately, librarians have been reticent to demand a voice in the development and design of information retrieval systems; so reticent, in fact, that there is little agreement even on how to describe the features each system possesses. Obviously, librarians need to be intelligent consumers of these systems, yet there are few satisfactory criteria against which to evaluate them.

One logical place to look for criteria for information system evaluation is the information retrieval research, but this research has often been isolated from the library context and virtually inaccessible to most practicing librarians. In the past, reference librarians have mediated the gap between the information retrieval machines—the large search services such as Dialog and BRS—and library users. Today, library users interact with information retrieval machines directly, chiefly through CD-ROMs and OPACs. The recent growth in end-user searching of all types has resulted in a literature characterized by laments about
the increased demand on the reference staff who feel called upon to instruct users individually or in classes, and by concerns that users are "not finding enough" or "not finding the best materials." But what is "enough?" And what are the "best materials?" These questions have usually been addressed in the context of reference service and mediated information retrieval, but when it comes to users' direct interaction with systems there is little information upon which to proceed.

Studies of end-user searching have focused on questions such as "Who is using the systems?" and "What are they finding?," or on management issues such as "How shall we select the best systems?" or "How shall we cope with the additional work load?" While there have been a few fine-grained analyses of the search experience of individual users, there have been even fewer studies that attempt to gauge users' success in fulfilling their actual information needs (Harter, 1990). Work done as prologue to expert system development has attempted to explicate the reference process in order to simulate and support reference tasks in an electronic environment. Also, some researchers are attempting to identify the core knowledge or expertise that should be incorporated into expert systems that could substitute for the assistance of a reference librarian in an information search (Fidel, 1986; Richardson, 1989). These are exciting and potentially productive research areas, but they are driven by a design perspective rather than an evaluation perspective. While it might be argued that until there are better information retrieval systems it is premature to be concerned with evaluation criteria, it is not too soon for librarians to articulate the criteria or goals of information retrieval systems. Furthermore, the design and development process is cyclical and iterative; what evaluation identifies as limitations in today's systems will lead to the innovations of tomorrow's systems.

These developments suggest that it would be useful and timely to look at the role of the user in evaluating the results of information retrieval. But in order to propose user-centered measures for information retrieval effectiveness, there must be a clear understanding of the goals of information retrieval so that appropriate evaluations can be performed. Some of the issues that must be addressed are:

- What are the implications of removing the intermediary from the information retrieval task?
- What does our knowledge of users' experience of information retrieval tell us about the goals of information search and retrieval, and how close we are to achieving them?
- How can the ways in which we ask our users about the services provided make the responses more useful?
USER, USE, AND USER-CENTERED STUDIES

User Studies

Most of the literature of the past three decades has focused on describing the characteristics of individuals and groups who use libraries or library information systems. Such studies answer questions like "Who is using the online catalog?," "Who are the users of MEDLINE CD-ROM?," and "Who are the end-users of Dialog?" They are generally descriptive, and examine variables such as profession, major, education, age, or sex. User surveys ask users to report their activities rather than directly observing their behavior. Little attention has been paid to defining what constituted a "use" and even less to understanding the nature of the interaction, and virtually no attention has been paid to non-users of libraries.

Use Studies

In the late 1970s, Brenda Dervin and Douglas Zweizig were some of the first to direct attention to the nature of users' interaction with libraries (Zweizig, 1977; Zweizig & Dervin, 1977). They found that information needs and uses were largely situation-bound and could not be generalized across all groups of users. While their work focused mostly on the use of libraries and information centers, other researchers, particularly in the 1980s, began to examine the process of searching (Markey, 1984; Kuhlthau, 1988). That is, they asked, "How and (sometimes) why is X system used?" "Was the search by author, subject, or title?" "Was the search for research, work, an assignment, or curiosity?" "How long was the search session?" "How many search statements were entered?" "How many modifications were made?" "What did the user do at the completion of the search?" Use studies often employ experimental designs or field research in which users are observed either directly or unobtrusively through transaction logs—the machine-readable record of the user's interaction with the computer (Nielsen, 1986). A recent book by David Bawden (1990) introduces a subcategory of use studies which he calls user-oriented evaluation. Bawden argues that in designing and testing information systems, one must move out of the laboratory and into the field, actually testing systems with real users. This may seem intuitively obvious, but unfortunately, it is often all too rarely done. Bawden also advocates the use of qualitative methods instead of or in addition to the experimental designs characteristic of information retrieval evaluations.
User-Centered Evaluation

User-centered evaluation goes one step beyond user-oriented evaluation. A user-centered study looks at the user in various settings—possibly not even library settings—to determine how the user behaves. The user-centered approach examines the information-seeking task in the context of human behavior in order to understand more completely the nature of user interaction with an information system. User-centered evaluation is based on the premise that understanding user behavior facilitates more effective system design and establishes criteria to use in evaluating the user's interaction with the system. These studies examine the user from a behavioral science perspective using methods common to psychology, sociology, and anthropology. While empirical methods such as experimentation are frequently employed, there has been an increased interest in qualitative methods that capture the complexity and diversity of human experience. In addition to observing behavior, a user-centered approach attempts to probe beneath the surface to get at subjective and affective factors.

Concern for the user and the context of information seeking and retrieval is not new, nor is it confined to library and information science. Donald Norman (1986) and Ben Shneiderman (1987) are well-known names in user-centered computer design. In library and information science, T. D. Wilson (1981) called for greater attention to the affective (or feeling) dimension of the user's situation nearly ten years ago. Wilson suggested that "qualitative research" leads to a "better understanding of the user" and "more effective information systems" (p. 11). For example, information may satisfy affective needs such as the need for security, for achievement, or for dominance. Qualitative methods are more appropriate to understanding the "humming, buzzing world" of the user than are the pure information science models derived from the communication theories of Shannon and Weaver (Shannon, 1948; Weaver, 1949).

The situations in which information is sought and used are social situations, where a whole host of factors—such as privacy or willingness to admit inadequacy and ask for help—impinge on the user and the information need. The context of the information-seeking task combined with the individual's personality structure, create affective states such as the need for achievement, and for self-expression and self-actualization (Wilson, 1981). Similarly, the subjective experience of the user can be examined in order to determine how it might be enhanced. For example, some studies have identified such affective dimensions of information retrieval as expectation, frustration, control, and fun (Dalrymple & Zweizig, 1990).
The user-centered approach, then, asks what the goals and needs of users are, what kind of tasks they wish to perform, and what methods they would prefer to use. Note that the user-centered approach starts with examining the user or the user's situation, and then goes about designing a system that will enable the user to achieve his or her goals. It does not start with the assumption that a certain objective amount of information is "appropriate" or "enough" for the task at hand. Having described the user-centered approach, the next section will summarize the history of evaluation in information retrieval and will describe the traditional criteria for retrieval effectiveness.

MEASURES OF EFFECTIVENESS IN INFORMATION RETRIEVAL

Precision and Recall

Ever since the Cranfield studies in the mid-1960s (Cleverdon, 1962; Cleverdon et al., 1966), the classic evaluative criteria of information retrieval system performance have been precision and recall, measures that were developed to evaluate the effectiveness of various types of indexing. Precision is defined as the proportion of documents retrieved that is relevant, while recall is defined as the proportion of the total relevant documents that is retrieved. These measures are expressed as a mathematical ratio, with precision generally inversely related to recall. That is, as recall increases, precision decreases, and vice versa. Despite their apparent simplicity, these are slippery concepts, depending for their definition on relevance judgements which are subjective at best. Because these criteria are document-based, they measure only the performance of the system in retrieving items predetermined to be "relevant" to the information need. They do not consider how the information will be used, or whether, in the judgment of the user, the documents fulfill the information need. These limitations of precision and recall have been acknowledged and the need for additional measures and different criteria for effectiveness has been identified. In addition to recognizing the limits of precision and recall, some of the basic assumptions underlying the study of information retrieval are being called into question by some information scientists (Winograd & Flores, 1987; Saracevic & Kantor, 1988). Thus, what appear at first to be objective quantitative measures depend, in part, on subjective judgments.

Relevance and Pertinence

We are seriously misled if we consider the relevant space of alternatives to be the space of all logical possibilities. Relevance always comes from a pre-orientation within a background. (Winograd & Flores, 1987, p. 149; emphasis added)
*Relevance* is defined as the degree of match between the search statement and the document retrieved. This is distinguished from *pertinence* in that the latter is defined as the degree to which the document retrieved matches the information need. Note that the difference between the two is the relationship between the search statement and the information need. Here is where the role of the intermediary comes in, and also the role of the system in helping the user to develop a search strategy. Research has shown that most users (indeed, even most searchers) have difficulty with search strategy.

One of the problems associated with precision and recall is the relevance judgement. Indeed, one of the first indications that there were cracks forming in the wall of precision and recall was Tefko Saracevic's (1975) review of relevance, in which he pointed out that relevance was a subjective and therefore unstable variable that was situation-dependent.

In a major study published recently, Paul Kantor and Saracevic (1988) presented findings that further questioned these traditional measures of retrieval effectiveness, particularly recall. They found that different searchers found different items in response to the same query. A similar phenomenon was identified by the author in a study of searching in both online and card catalogs (Dalrymple, 1990).

Precision and recall need not be discarded as evaluative measures; they remain useful concepts, but they must be interpreted cautiously in terms of a variety of other factors. For example, when determining precision, is the user required to actually examine the documents that the citations refer to? If so, then another variable is being tested: the accuracy of indexing. If not, then what is being measured is the degree of fit between the user's search statement as entered into the system and the indexing terms assigned to the documents. The "fit" between the documents and the user's information need is not being considered. After all, it is the skill of the indexer in representing the contents of the document that is tested when the user compares the retrieved document to the original information need; the retrieved citation is merely an intermediary step. In fact, the Cranfield studies themselves were designed to do just that—test the accuracy of indexing, not evaluate the "success" or "value" of the information retrieval system or service.

If users are not required to examine the documents in order to make relevance judgements, then what shall be substituted? Users make evaluations simply on the retrieved citation. Brian Haynes (1990) found that more than half (60 percent) of the physicians observed made clinical decisions based on abstracts and citations retrieved from MEDLINE without actually examining the documents. Beth Sandore (1990) found in a recent study of a large Illinois public library that users employ various strategies in determining relevancy of retrieved items—"the most
common appear to be arbitrary choice or cursory review” (p. 52). Several issues can be raised immediately. First, without evaluation studies in which users actually examine the documents—i.e., read the articles and absorb the information—then perhaps what is being evaluated is the ability of a bibliographic citation or abstract to catch the user's attention and to convey information. Second, how do relevance judgments change when users read the documents? Third, what other factors affect the user's selection of citations from a retrieved list?

Recall has also come under scrutiny as an effectiveness measure. Since it is virtually impossible to determine the proportion of relevant items in an information system except in a controlled laboratory study, it may be more useful to regard recall as approximating the answer to the question, “How much is enough?” Sandore found that “many patrons use—that is, follow up and obtain the document—much less information than they actually receive” (p. 51). In her provocatively titled article, “The Fallacy of the Perfect 30-Item Search,” Marcia Bates (1984) grappled with the notion of an ideal retrieval set size, but these studies have focused on mediated information services. Little has been done to examine how much is enough for users when they access information systems directly. Stephen Wiberley and Robert Daugherty (1988) suggest that the optimum number of references for users may differ depending on whether they receive a printed bibliography from a mediated search (50) or search a system directly such as an OPAC (35). Although one limitation to recall as a measure is that it requires users to describe what they don't know or to estimate the magnitude of what might be missing, perhaps a more serious limitation is that it is not sensitive to the ever-increasing threat of information overload. As systems increase in size, users are more likely to receive too much rather than not enough; when retrieved documents are presented in reverse chronological order (as is the case in virtually all information retrieval systems), users may find themselves restricted to seeing only the most recent, rather than the most useful, items.

Other Measures of Information Retrieval Effectiveness

In addition to precision and recall, there are other evaluative measures that have enjoyed a long history in information retrieval research. Some of these dimensions are cost (in money, time, and labor), novelty, and satisfaction related to information need.

Cost

Cost of online retrieval is subject to external pressures of the marketplace. For example, in 1990, current pricing algorithms of major vendors were changing away from connect time charge and toward use
charges, which may have the effect of reducing the incentive to create highly efficient searches. Access to optical disk systems, online catalogs, and local databases provided directly to the user with neither connect charges nor use charges creates an incentive toward greater use regardless of the efficiency of the search strategy or the size of the retrieval set.

F. W. Lancaster (1977) observed that precision can also be treated as a cost in that it is an indirect measure of the time and effort expended to refine a search and review results (p. 144-46). In direct access systems, precision may be achieved iteratively, much more so than with delegated searches. The user can decide where the effort is going to be expended—in doing a tutorial, in learning to be a so-called “power user,” or in doggedly going through large retrieval sets.

**Novelty**

Novelty is defined as the proportion of the retrieved items not already known to the user (Lancaster, 1979, pp. 132-33). With mediated searches, novelty is usually measured by asking the user to indicate which of the items retrieved were previously known. Novelty, of course, is related to the degree of subject expertise possessed by the user. That is, a subject specialist is quite likely to be familiar with a great many of the items retrieved in an area of expertise; the only items that are truly novel are those recently published. For the subject specialist, presenting the most recent items first makes sense; but this design decision may not apply to all, or even most, users in nonspecialized libraries. For those users, it may make much more sense to present the most relevant items first; this can be done by assigning mathematical weights based on term frequency or location. Such systems currently exist on a small scale, but are not yet widely available. Regardless of which model is chosen (and ideally, both options should be available in any given system to accommodate various knowledge states in users), the point is that both approaches recognize that the effectiveness of the retrieval is affected by the user situation.

**Information Need**

In order to discuss satisfaction it is necessary to address the problem of information need. Some researchers sidestep the problematic area of information need, arguing that because these problems are abstract, unobservable, and subject to change, it is futile to include them in research and evaluation. Others, while admitting these problems, nevertheless call for increased efforts in trying to grapple with them. One of the most convincing statements of the importance of understanding information needs was made by Brenda Dervin and Michael Nilan (1986) in a review of information needs and uses. They call for a paradigm shift that:
posits information as something constructed by human beings....It focuses on understanding information use in particular situations and is concerned with what leads up to and what follows intersections with systems. It focuses on the users. It examines the system only as seen by the user. It asks many "how" questions—e.g., how do people define needs in different situations, how do they present these needs to systems, and how do they make use of what system offer them. (p. 16)

Within this paradigm, information needs focus on "what is missing for users (i.e., what gaps they face)" (p. 17) rather than on what the information system possesses.

Focusing on the user's information need may lead to a reconsideration of the assumptions underlying library and information systems and services. As an example, consider Karen Markey's (1984) research in online catalogs. By observing what users actually do when searching an online catalog, she discovered that a remarkable number of catalog users were conducting subject or topical searches in the catalog, rather than known-item searches. Her findings prompted a reconsideration of how libraries approach the study of catalogs, and even how they approach their evaluation and improvement. Catalogs are now seen as subject access mechanisms, and there have been many proposals as to how to go about improving subject access in online catalogs. Valuable as this research is, it has proceeded without a thorough examination of librarians' assumptions about the function of the catalog. That is, there has been no attempt to ascertain what users need the catalog for, what their purposes are in searching the catalog, what they expect to find, what need prompts them to approach the catalog—or even the library, for that matter—and how and whether it meets those needs. Until these questions are asked and answers attempted, librarians shall be bound within the old paradigm that defines an information need as something that can be satisfied by what is available in information systems.

USER-CENTERED MEASURES OF INFORMATION RETRIEVAL

Satisfaction

....satisfaction is determined not by the world but by a declaration on the part of the requestor that a condition is satisfied. (Winograd & Flores, 1987, p. 171)

It has been suggested that the satisfaction of a human user rather than the objective analysis of the technological power of a particular system may be a criterion for evaluation. This is generally not the position that has been taken by library and information researchers,
but the literature is by no means devoid of concern for user satisfaction. When one reviews two decades of library and information science research, a renewed interest in affective measures seems to be on the horizon. The waxing and waning of interest in affective measures in information retrieval may parallel the changing role of the intermediary in information retrieval. That is, affective measures have been attributed to the "human touch" in information service rather than to the machines that perform the information retrieval task.

The user's satisfaction with the outcome of the search when it is performed by an intermediary was investigated by Judith Tessier, Wayne Crouch and Pauline Atherton (1977). Carol Fenichel (1980) used both a semantic differential and a five-point rating scale to measure intermediaries' satisfaction with their own searches and found no evidence to support the contention that intermediary searchers are good evaluators of their searches. Sandore (1990) found that there was very little association between search satisfaction and search results as indicated by precision; patrons who were dissatisfied with the results still reported satisfaction with the service. In both of these studies, satisfaction with the search experience is separated from satisfaction with the retrieved results as measured by precision. Satisfaction is indeed a complex notion that may be affected by the point in time at which the measure is taken; it can be affected by the items that the user selects, the difficulty encountered in locating the documents, and the information contained in the documents.

Considering the context of the information retrieval experience, particularly for end-users, underscores both the importance and the multidimensionality of affective—that is, feeling—measures. Judith Tessier (1977) identified four distinct aspects of satisfaction with the information retrieval process: output, interaction with intermediary, service policies, and the library as a whole. She wrote: "Satisfaction is clearly a state of mind experienced (or not experienced) by the user...a state experienced inside the user's head..." (p. 383) that is both intellectual and emotional. She observed that the user's satisfaction is a function of how well the product fits his or her requirement (or need), that satisfaction is experienced in the framework of expectations, and that people seek a solution within an acceptable range rather than an ideal or perfect solution.

Tessier's work is insightful, but it has rarely been integrated into studies of end-user searching in today's environment. In most studies of end-user searching, satisfaction is treated as unidimensional: users are either satisfied or they are not. Furthermore, most studies depend on users' self-assessments, and most users are not adequately informed about the system's capabilities. Users have notoriously low expectations and are usually unimaginative in identifying additional features that
would be desirable, nor are they presented with alternatives from which to select. While retaining a degree of skepticism when users respond on a questionnaire that they are "satisfied," it must be acknowledged that it is the users themselves that determine their response to systems. And while it would be desirable for users to be more discriminating, little has been done to provide alternatives or even simply to ask users to rank various features of a system or its output. Users are not asked, "Did the information make a difference?" or better yet, "How did it make a difference?" In general, users have not been asked to describe their experiences in any but the simplest terms.

Much of the interest in examining user responses that was begun in the 1970s, when systems were first made available for direct access, waned over the past two decades when most searching was done by intermediaries. Stimulated by the current interest in end-user searching, it is interesting to return to some of the approaches used twenty years ago. For example, Jeffrey Katzer (1972) used factor analysis with a semantic differential to identify three dimensions that were relevant to information retrieval systems: the evaluation of the system (slow-fast, active-passive, valuable-worthless), the desirability of the system (kind-cruel, beautiful-ugly, friendly-unfriendly), and the enormity of the system (complex-simple, big-small).

The author and Douglas L. Zweizig recently factor-analyzed data from a questionnaire designed to determine users' satisfaction with the catalog search process (Dalrymple & Zweizig, 1990). The data were collected at the conclusion of experimental search sessions in which users were randomly assigned to perform topical searches in either a card catalog or an online catalog. Interestingly, the objective measures of catalog performance failed to discriminate between the two catalogs' conditions, and simple descriptive comparisons of the two groups did not reflect differences, either. But when the questionnaire data were subjected to a factor analysis, two primary factors were identified: Benefits and Frustration. Frustration emerged from responses such as "it was difficult to find the right words, it was frustrating, and confusing to search" (p. 22). Additional factors were also identified, and the strength of each of the factors differed depending on the catalog setting—card or online—and the way in which these factors correlated with other aspects of the search differed, depending on the type of catalog. For example, in the OPAC, users who reformulated their searches often, scored high on the Benefits factor, but in the card catalog, the reverse was true. Intuitively, it makes sense that changing direction in an online search is easier than having to relocate into another section of the card catalog. Thus, in the card catalog, redirecting a search (reformulating) is perceived as frustrating and detracts from the user's perceived benefits, but reformulation is a natural part of the search activity in the OPAC.
and so correlates positively with the Benefits factor. Also, users were asked to assess the results they achieved on their searches. Subjects who enjoyed their experience searching in the OPAC viewed their results favorably, while in the card catalog, users viewed their search results favorably despite the frustration they experienced.

These examples indicate the complexity and multidimensional nature of affective measures, and show that they are sensitive to a variety of situational factors. In the next section, context as a factor in evaluating the impact of information retrieval will be discussed.

**Context and Impact**

Reference librarians are well aware of the importance of understanding the context of an information request, and the literature of the reference interview is replete with discussions of symbolic and nonverbal aspects of the communication between reference librarian and user. Much less attention has been paid to contextual aspects of end-user searching of electronic information systems, by either librarians or information scientists. Two studies (Saracevic & Kantor, 1988; Dalrymple, 1990) examined the sets of items retrieved by individual searchers and found that the overlap was relatively low, even though the databases searched were identical. That is, given the same questions, different searchers tended to select a few terms that were the same and a considerably larger number that were different. This finding held true both for experienced intermediaries and for end-users in both database searches and OPAC searches. In explaining these differences, both studies acknowledged the importance of the user's context in determining the direction of the search.

Because context is such a powerful element in retrieval effectiveness, looking only at “objective” measures such as retrieval sets and transaction log data may have limited usefulness in determining retrieval effectiveness. Rather, it may be better to look at human beings and the situations in which they find themselves, and to evaluate retrieval effectiveness in terms of the user's context (Dervin & Nilan, 1986).

Not only does context affect retrieval, but it also affects the progress of the search through system feedback. The psychological aspects of information retrieval are receiving a great deal of attention by information scientists, computer scientists, and cognitive scientists alike. Studies of computerized searches can often reveal much about the ways in which individuals interpret queries, pose questions, select terms, and understand and evaluate information. One might even say that the information search provides a kind of laboratory for understanding human information processing. By examining in detail the history of a search, both from the system’s perspective (through the transaction
log) and from the user's perspective (through "talking aloud" and in-depth interviews), insight can be gained into the factors that affect the search, and these can be used to articulate the criteria against which information systems will be evaluated.

Some of the models used to design information systems underscore the role of psychological understanding of the search process. One is a communication model in which information retrieval is seen as a conversation between user and information system; another is a memory model in which information retrieval is seen as analogous to retrieval from human long-term memory. In the conversational model, the user and the system engage in a "dialogue" in which each "participant" attempts to gain an understanding of the other. For example, an expert system embedded in an information retrieval system might prompt the user to provide more specific information about what is needed (Do you want books or articles?), to provide synonyms (What do you mean?), or to limit the retrieval in some way (Do you want materials only in English? Only in the last five years? Only available in this library?). By answering the questions and engaging in the dialogue, the user participates in the process.

In retrieving from long-term memory, the searcher is even more active. In this model, the user finds a context by entering terms into a file and displaying the results until the context that seems most likely to meet the information need is found. The user searches that context for other similar items until all probable useful items are found, and then "verifies" them by asking, "Will these meet my need? Is this what I am looking for? Does this make sense?" In both models, the user performs the evaluative judgment based on her or his situation in the world. Regardless of the particular model chosen, the point is that both models are iterative and interactive. That is, they assume that the user is an active participant in the information retrieval process, and that continuous feedback from both system and user, one to the other, enables the process to advance and to continually improve.

But how does this fit into evaluation of information retrieval systems and services in a library? Stepping back for just a moment, it is essential to ask what it is that information retrieval systems are designed to do. For example, should catalogs do as Patrick Wilson (1983) suggests and simply verify the existence of an item in a collection? Or shall they act as knowledge banks, capable of providing information that goes well beyond simply indicating probable shelf locations for relevant items? Shall databases provide "quality-filtered" information that can support decision-making in highly specific areas, or shall they simply indicate the existence of an article on a topic? Shall systems "stand
in" for reference librarians, and if so, is it reasonable to use the same criteria in evaluating an information system as in evaluating reference personnel?

Definitive answers to these questions do not yet exist, nor will one set of answers apply to all systems, to all libraries, and to all users, all of the time. By placing users and their needs much closer to the center of evaluation, methodologies can be employed that are sensitive to situations and contexts of users. "Qualitative evaluation tells us how well we have met the patron's needs" (Westbrook, 1990, p. 73).

Exactly how one should begin to both answer and ask these questions suggests a methodological discussion. Increasingly, researchers in user studies call for applying qualitative methods—that is, in-depth investigations often using case study, which seek to study the behavior of individuals in all of the complexity of their real-life situations. Qualitative evaluation seeks to improve systems and services through a cyclical process, in which both quantitative (statistical) and qualitative methods are employed, each used to check and illuminate the other. Some methods such as observation and interviews are particularly well-suited to field studies to which librarians can contribute substantially. Gathering the data in qualitative studies is done over time, often by participant observers who possess a knowledge of the setting and who could be expected to have insight into the situation. While simply "being on the scene" is hardly enough to qualify one as a researcher/evaluator, cooperative research and evaluation projects in which librarians play a significant role can do much to enhance one's understanding of the issues and problems associated with satisfying information needs. What follows is a discussion of some of the dimensions of the user's experience with an assessment of information retrieval.

Although Bawden's work presents it, it is necessary to go one step further—to question librarianship's assumptions about users and the purpose of information retrieval, and then to move to an in-depth exploration of what it means to seek information in libraries today. Until answers to such questions as "What are the user's expectations for how a system functions?", "What needs does it meet?", and "What is the experience of searching really like for the user?" are found, criteria for evaluating retrieval effectiveness will not be improved.

CONCLUSION

...the involvement of the practitioner is a sine qua non for the success of user-oriented evaluation. (Bawden, 1990, p. 101)

Information retrieval has been locked into a rationalistic, empirical framework which is no longer adequate. A different framework of
analysis, design, and evaluation that is contextual in nature is needed; such a framework is both interpretive and phenomenological. It implies that information retrieval tasks are embedded in everyday life, and that meanings arise from individuals and from situations and are not generalizable except in a very limited sense. Users are diverse, and their situations are diverse as well. Their needs differ depending on their situation in time and space.

Information systems may therefore differ, offering diverse capabilities—often simultaneously within the same system—which provide an array of options the user can select. For example, such systems may offer interfaces tailored to many skill and knowledge levels; they may allow users to customize their access by adding their own entry vocabularies or remembering preferred search parameters; or they may provide a variety of output and display options. In order to move beyond the present-day large, rather brittle systems which are designed to be evaluated on precision and recall, evaluation studies must be conducted that can be used in the design of new systems. By focusing on users as the basis for evaluative criteria, new systems that are more responsive and adaptive to diverse situations can be created.

User-centered criteria—affective measures such as user satisfaction and situational factors such as context—are beginning to be used in research and evaluation. But this is just a beginning. Librarians and researchers alike must retain and refine their powers of critical observation about user behavior and attempt to look at both the antecedents and the results of information retrieval.

The methods used to gain insight into these issues are frequently case studies, focus groups, or in-depth interviews which, when combined with objective measures, can afford effective methods of research and evaluation. When placing the user at the center of evaluations, it is important not to take behaviors at face value but to probe beneath the surface. In order to do this successfully, it can mean small scale, in-depth studies carried out by astute, thoughtful individuals—ideally, a combination of both practitioners and researchers.

REFERENCES

User-Centered Evaluation of Information Retrieval


