Reflections on Information Retrieval Evaluation

Mei-Mei Wu & Diane H. Sonnenwald,
National Taiwan Normal University / University of North Carolina at Chapel Hill, Taiwan / USA

Abstract

Information retrieval (IR) research has primarily consisted of two paradigms: systems-oriented research and user studies. Systems-oriented research includes IR algorithm development and evaluation and, to some degree, human-system interaction. User studies include human information behavior and information seeking research. These paradigms have contributed new IR algorithms and insights into IR processes. However, problematic issues concerning IR evaluation exist in both communities. For example, research methods, measures, and metrics that can yield generalizable results are problematic for systems-oriented research and user studies. We propose that establishing a dialog between these communities and synthesizing their methods may increase the effectiveness of both. In this paper, a brief overview of IR evaluation, including systems-oriented research and user studies, is presented and problematic issues in IR evaluation are discussed. We synthesize these approaches and propose a framework for evaluation that is based on attributes that have been shown to influence adoption of innovations (Rogers, 1995).

1. Introduction

With the continued information explosion, including the emergence of the internet and digital library initiatives, information retrieval (IR) evaluation has increased in importance and is an active area of research and development. For example, research funding agencies now more than ever require IR research, including digital library projects, to illustrate their applicability and utility to real world problems. This requires evaluation. Furthermore, as advanced information retrieval systems move from research to the real world of commercial competition, designers and developers, vendors, and sales representatives of new information products, such as electronic (or digital) books, search engines, and personal Internet filters want to know whether their products offer potential users and purchasers competitive advantages. In addition, other research has shown that individuals prefer asking people for information, and only come to use IR systems when nothing else is available (e.g., Chatman, 1996; Kuhlthau, 1993; Sonnenwald, 1999).

There are two major paradigms of IR research today: the "systems-oriented" and "user studies" paradigms. This paper reviews major issues of IR evaluation in the systems and user paradigms, and discusses the strengths and weaknesses in both paradigms. There has been little dialog between people adopting these two paradigms (Saracevic, 1997). In this paper, the paradigms are synthesized in a new, proposed framework for IR evaluation. The proposed framework is based on attributes that have been shown to influence adoption of innovations (Rogers, 1995.) Criteria and measures for these attributes are suggested.

2. Overview of Systems-oriented IR Evaluation

2.1 Characteristics and Contributions

Systems-oriented IR evaluation dates back to the 1950s, when the first Cranfield tests were launched in England (Sparck Jones, 1981). The Cranfield tests used an experimental design setting, the purposes of which were to examine whether different indexing functions performed differently. The goal was to identify which algorithm performed best according to standard measures and metrics. The systems-oriented experimental research tradition was thus founded. Salton in the United States was the first to extend the experimental approach to include evaluation of vector space IR algorithms (Salton & McGill, 1983).

This experimental approach includes several major elements: a set of "documents" (i.e., a representation of documents that often includes title and author information and abstracts and only recently has begun to include whole documents), IR processes (that include query processing and the IR algorithm or indexing method to be tested), a set of artificial or semi-artificial queries (which are created by skilled intermediaries or modified from end-users' search requests), and a set of answers (the relevant documents as pre-determined by domain experts). The queries are input to the IR process, and the output consists of a list of documents identified as relevant by the process. This list is compared to a list of relevant documents pre-determined independently by domain experts. This comparison is made using two measures, recall and precision. "Recall" measures the proportion of relevant documents retrieved, and is calculated as the total number of relevant documents identified by the IR process divided by the total number of relevant documents previously identified (by domain experts). "Precision" is the proportion of retrieved documents that are actually relevant (as previously determined.) The queries, document collection, and list of relevant documents are frequently "standard" in the sense that many researchers share these components. This sharing has enabled them to compare evaluation results.

Benefits of this approach to IR evaluation include the development of advanced, performance IR algorithms including vector space and probabilistic retrieval algorithms, relevance feedback algorithms, and query processing algorithms. These results have recently been incorporated into search engines that facilitate searching large and varied collections of web pages on the Internet, and into online public access catalogues that facilitate searching large library collections.

Before 1990, this type of IR evaluation research was primarily carried out by individuals and small groups of researchers at a variety of locations throughout the world. This meant that test collections, including queries, documents, and relevance judgments by experts were relatively small in scope and few in number because individuals did not have the available resources required to build extensive test collections. To address these issues, the National Institute of Standards and Technology (NIST) and the Defense Advance Research Project Agency (DARPA) in the United States launched the Text REtrieval Conference (TREC) in 1991.

TREC has four major goals: to encourage research in text retrieval based on large-scale test collections; to increase communication among researchers in industry, academia, and government; to speed the transfer of technology from research laboratories into commercial products; and to increase the availability of appropriate evaluation techniques (Harmon, 1996). Yearly, individual researchers and/or research teams participate in TREC.

TREC consists of a document set, topic (or query) sets, relevance judgments provided by domain experts, and tasks and tracks. The document collection is provided without cost to everyone who participates. It consists of heterogeneous collections from several sources such as the Financial times Limited, US Congressional Record of the 103rd Congress (1993), Foreign Broadcast Information Services (1996), and Los Angeles Times (1989, 1990) [1]. Documents are primarily full text and vary in number and length; together they comprise five gigabytes of data. "Topics" used in TREC are intentionally designed to represent user information need statements, and are used by all participants. TREC topics were gathered by skilled intermediaries and include a list of important concepts and a more or less precise statement of criteria for determining relevance for the judges. The relevance judgments concerning documents in the document collection are pooled from the participating research groups and 3 domain experts. The top 100 relevant documents from each system are merged into a list, which is then evaluated by domain experts. Their judgments are pooled to yield the final overall relevance judgments for the topics and documents.

[1] For a complete list of sources see NIST, 1999.

In TREC, new types of IR tests, called tracks, are proposed each year, and each track is designed to address a different IR problem. Researchers or research groups may select which track(s) they wish to participate in. Currently, in TREC 8, the tracks include (TREC, 1999):

(a) Cross-language track: Documents are in English, German, French or Italian, and topics (or queries) are in each language. The challenge is to retrieve relevant documents regardless of language.

(b) Filtering track: The topics are stable and some relevant documents are known. The system must make a binary decision as to whether new, incoming documents are relevant to the topics.

(c) Interactive track: This track is used to study user interaction with IR systems, e.g., the role of relevance feedback in IR.

(d) Query track: This track examines the effects of query variability and analysis on retrieval performance.

(e) Question answering track: This is a new track that strives to address "information" retrieval as compared to "document" retrieval. For a set of 200 questions, systems must produce a text extract that answers the questions. The text extracts should range from short phrases (2-3 words) to an entire document (1,000 words.)

(f) Spoken document retrieval track: This track investigates systems' ability to retrieve spoken documents (recordings of speech.)

(g) Web track: In this track, the document set is 2 gigabytes of web data. This track will encourage researchers to investigate whether links can be used to enhance IR.

Contributions from TREC include the development of a community of researchers who compete and compare their algorithms on a regular basis using large document collections. This has led to refined algorithms that produce improve recall and precision measurements, and work well with very large collections of documents. It has also broadened IR evaluation research to include cross-language IR, spoken document retrieval, and the investigation of relevance feedback.

The TREC approach is essential to improve the technical processing level, e.g., language processing and algorithm design, and is being replicated internationally. For example, Japan has initiated a similar research endeavor. A test collection for Japanese IR systems has been developed by Information Processing Society of Japan (IPSJ) and Real World Computing Project (RWCP). Their initiative program will begin in 1999. The working group include members from: ERI, University of Toyko, Fuji Xerox, Fujitsu Laboratories, Keio University, Mitsubish Electric, NACSIS, NEC, NTT, NTT Data, Ricoh, SHARP, Tokyo Institute of Technology, Toshiba, ULIS. As the population of Chinese language of Internet users increases, it is particularly important to promote Chinese IR research. Wu (1998) has proposed a prototype for CTREC, a Chinese language TREC initiative. It takes a group effort to construct such an environment that can facilitate IR evaluation research.

2.2 Issues

This approach to evaluation, however, yields some problems and criticisms (e.g., Ellis, 1996; Harter & Hert, 1997; Saracevic, 1995; Tague-Sutcliffe, 1996b). Harter & Hert (1997) classified evaluation problems with the systems-oriented approach into four categories: validity and reliability, generalizability, usefulness, and conceptual. Issues regarding validity and reliability include the omission or the passive role of the user, and the laboratory setting versus real world settings which includes a variety of dynamic human information-seeking behaviors, human-system interaction, IR system and other computer-based and non-computer-based system interactions. In addition, there are untested assumptions regarding relevance judgments, e.g., the assumption that relevance judgments for a number of documents are, for any given search question, independent of one another and are not subjective and dependent on a given time and place.

In terms of generalizability, it is argued that there has been lack of random sampling of users and information needs so that queries used in the experiments are perhaps not representative and that experimental test collections are small and deal primarily with science and technology topics. These issues are common in research, and we see them also in user studies. Another issue is that findings appear to conflict with common sense and experience. For example, Belkin et al (1996) found no relationship between any demographic, or experience variables, and IR system performance. One would expect that expertise in IR in general and with the IR system under evaluation would have an impact on system performance.

With respect to usefulness, criticisms include: lack of applicability to operational systems, which means that in general, operational systems must be evaluated using other criteria; and

weak power of findings, i.e., weak in explanation, prediction, and control of the phenomenon under investigation. For example, Su (1992) found precision was not significantly correlated with users' perceptions of success. Instead, Su found that users' satisfaction with completeness of search results and value of search results as a whole, among other measures, were significantly correlated with success.

Conceptually, there is limited theoretical support for the measures and metrics used in this approach to IR evaluation. It is not clear that recall and precision are meaningful to individuals seeking information, or that statistical differences in these measures between systems is significant in real world contexts and situations. However, TREC does include an interactive track that focuses on user interaction with IR systems. This track is an attempt to bridge the gap between the systems-oriented approach and user studies.

2. User Studies

3.1 Characteristics and Contributions

User studies in IR evaluation began in the 1970s when several commercial IR systems became available. Saracevic and Kantor's study at Case Western University in the early 1980s was a pioneering study (Saracevic & Kantor, 1988a; 1988b). The main purposes of the study were to identify the searcher's searching behavior and user satisfaction, and to determine system effectiveness. Comparing with the experimental evaluation approach described earlier, this study focused on users; user satisfaction was equated with system effectiveness. Whether user information needs were being met, whether information retrieved was useful or not, and whether the human-computer interface or interaction was "user friendly" were the major questions of early user studies in IR evaluation.

User studies have evolved to include research on human information behavior which can be broadly defined as the processes and outcomes of information exploration, seeking, filtering, use, provisioning, and dissemination. It focuses on "real" users, such as children (.e.g., Solomon,1994; Borgman, Hirsh, & Gallagher, 1995), high school students (e.g., Kuhlthau, 1993), project team members in a variety of work organizations (e.g., Sonnenwald & Pejtersen, 1994; Algon, 1996; Solomon, 1997; Kuhlthau, 1996), clients of public health services, and library professionals (e.g., Iivonen, 1995; Iivonen & Sonnenwald, 1997). These studies typically use more qualitative methods and measures than systems-oriented research. Methods used include interviews, observations, think-aloud experiments, and surveys to determine users' perspectives and behavior with respect to information retrieval. Because the focus is to build theories and models of human information behavior that can

lead to new designs of IR systems and services, IR systems may not be explicitly included or evaluated in these studies.

The international conference, Information Seeking in Context (Vakkari, Savolainen, & Dervin, 1997; Wilson, 1999), has emerged as a forum for researchers in these area to explore methods and research results. Another emerging forum is the newly formed special interest group (SIG) on information needs, seeking and use of the American Society of Information Science (ASIS) (Cheng & Shaw, 1999). These forums are analogous to TREC in that they strive to encourage research in user studies; to increase communication among researchers in industry, academia, and government; and to identify and refine appropriate research techniques. However, they are dissimilar in that a standard evaluation methodology is not promoted.

User studies have made a variety of contributions to the field of information retrieval. These contributions include the identification of human information seeking behavior, bridging the gap between individuals' information needs and IR systems and leading to new types of IR systems that include graphical human-computer interfaces (e.g., Pejtersen, 1989; Borgman, Hirsch, & Gallagher, 1995), new types of information about information resources that aid the user in finding relevant resources, and the need to include new types of information resources in IR systems (e.g., Sonnenwald et al, 1999).

In addition, user studies have helped uncover the dynamic and situational nature of relevance, which has led to the reconsideration of IR evaluation criteria. For example, user studies had illustrated that the concept of relevance is continuous and not dichotomous, and, that relevance judgments are subjective and situational but not objective or logical. Schamber (1994) sees relevance, as reflected in the user studies literature, as subjective, cognitive, situational, psychological, multidimensional, dynamic, and measurable. Schamber lists 80 factors in six broad categories that constitute a partial list of variables that have been found to affect, or have been suggested to affect, relevance. These categories are: judgments, requests, document, information systems, judgment conditions, and choice of scale.

Saracevic (1996) pointed out that relevance indicates a relation, and different manifestations of relevance encompass different relations. The five types of manifestations of relevance are: system or algorithmic relevance, topical or subject relevance, cognitive relevance or pertinence, situational relevance or utility, and finally motivational or affective relevance.

System or algorithmic relevance refers to the relation between a query and information objects, in the file of a system as retrieved, or as failed to be retrieved, by a given procedure

or algorithm, comparative effectiveness is inferring relevance is the criterion for system relevance. Topical or subject reference refers to the relation between the subject or topic expressed in a query, and topic or subject covered by retrieved texts, or more broadly, by texts in the systems file, or even in existence. Aboutness is the criterion by which topicality is inferred. Cognitive relevance or pertinence refers to the relation between the state of knowledge and cognitive information need of a user, and texts retrieved, or in the file of a system, or even in existence. Cognitive correspondence, informativeness, novelty, information quality, and the like are criteria by which cognitive relevance is inferred. Situational relevance or utility refers to the relation between the situation, task, or problem at hand, and texts retrieved by a system or in the file of a system, or even in existence. Usefulness in decision making, appropriateness of information in resolution of a problem, reduction of uncertainty, and the like are criteria by which situational relevance is inferred. Motivational or affective relevance refers to the relation between the intents, goals, and motivations of a user, and texts retrieved by a system or in the file of a system, or even in existence. Satisfaction, success, accomplishment, and the like are criteria for inferring motivational relevance.

In her recent research, Wu (1998) suggest that relevance judged by the intermediaries and by the end-users appear to reflect these different notions of relevance. For example she found there was no statistically significance between relevance judgments made by three professional intermediaries, however, there was a slightly difference between the end-users' relevance judgment and the intermediaries. When relevance is seen from the different perspectives of cognitive, situational, or motivational, it is difficult to use it as a criterion to compare the effectiveness or success across different IR systems. This problem can not be solved unless the same group of study participants can be tested across systems or different evaluation measures are discovered.

3.2 Issues

Similar to the systems-oriented approach in IR, issues concerning the generalizability, utility and conceptualization of user studies have emerged. Because user studies typically focus on a specific user population and that population may be small in number, their results may be contingent on person, place and time and not generalizable to larger or different populations. To increase their generalizability, studies need to be replicated in a variety of settings. However, this is problematic because user studies are often time intensive; they can take months and years to complete. Several studies have been replicated or are being replicated (e.g., Kuhlthau, 1993; Sonnenwald, 1996, 1997; Spink, Wilson, Ellis & Ford, 1998), and more work along these lines needs to be done.

A second issue focuses on utility. It may be difficult to see the relationship between the results of user studies and IR system design and evaluation. The question is, how can results of user studies inform IR system design? Often, this is a practical as well as theoretical issue. Researchers who conduct user studies are experts in social and behavioral science research methods and theories, not necessarily in technology. They themselves do not know how to build systems based on their research results, or how to frame or translate their results so others can. Indeed, collaboration among researchers in user studies and IR systems is rare. How can this gap be bridged?

Conceptually, it can be difficult to compare or synthesize results from user studies because different data collection and analysis methods and different levels of analysis and conceptualization are done across studies. When different levels of analysis and different conceptual constructs are built across studies, it is difficult to synthesize and/or compare the results. Recent research (e.g., Ingwersen, 1997; Saracevic, 1995; Wilson, 1997; Sonnenwald, 1999) has begun to attempt this and it will be interesting to see how these and other efforts unfold in the future. Theories of human information behavior that span across contexts and situations are needed in order to identify new measures and evaluation methods for IR. The theories are just beginning to emerge; we will need to wait to see if new measures will emerge as well.

2. Bridging Different Paradigms

4.1 The Need to Bridge Paradigms

Saracevic (1995) has proposed a series of questions regarding IR evaluation:

- How successful was and is information retrieved in resolving the problem of information explosion in the areas applied?
- How well does IR support people in situations when they are confronted with problems of seeking, finding, using, and interacting with information from the mass of existing information and myriad of choices available?
- How does all this information, and associated information technology and information systems effect our work, leisure, society, and culture?
- How do IR and related applications reorder life?

These questions are particularly vital in the context of today's society when increasingly more information is being represented in digital form and searching for information using techniques, such as physically browsing and searching library stacks, that have been developed and refined over hundreds of years will no longer be available. The complicated,

digital surrogates could make retrieval tasks even more complex (Saracevic, 1995; Tague-Sutcliffe, 1996b; Ellis, 1996.)

These issues can not be answered by systems-oriented or user studies alone. There is a need to synthesize the systems-oriented laboratory research approach and real world situations, contexts, individuals and organizations in order to demonstrate utility and effectiveness from users' perspectives.

4.2 Related frameworks

Several evaluation frameworks have merged from related disciplines such as human factors and cognitive science. One such framework is usability engineering (e.g., Whiteside, Bennett & Holtzblatt, 1988) that provides methods and techniques for evaluating the usability of the human-computer interface in relationship to tasks users are expected to perform with the system. Other frameworks, such as cognitive systems engineering (e.g., Rasmussen, Pejtersen & Goodstein, 1994) suggest that dimensions of cognitive work, including the work domain problem space, tasks, mental strategies, work roles, and management culture, need to be considered and evaluated. Socio-technical design (e.g., Mumford, 1994) suggests that the system and its impact on its work and organizational context also needs to be considered and evaluated. Others (e.g., Beer, 1976) suggest that the larger societal impact also needs to be considered and evaluated.

In the case of digital libraries, Belkin (1994) suggests four foci for evaluation. These are a focus on goals, problems, tasks that lead users to engage in interaction with information objects; focus on access, rather than on institution; focus on interactions with information objects; and a focus on impact on users and on user communities. Possible evaluation criteria include goals met, effectiveness, use and usability, and utility. Methods, such as observation and experimentation with interaction in digital library contexts, are suggested.

Saracevic (1995) proposes six levels of IR evaluation: engineering, input, processing, output, use and user, and social. On the engineering level, questions concerning hardware and software performance are addressed. Thus, computational effectiveness and efficiency of given retrieval methods and algorithms are investigated. On the input level, questions about the inputs to and contents of the system are investigated. Thus, questions about coverage in the designated area are asked. On the processing level, questions about the way the inputs are processed arise. Thus the assessment of performance of algorithms, techniques, approaches, and the like are acknowledged. On the output level, questions about interactions with the system and obtained output(s) are addressed. Thus, the evaluation

criteria include assessment of searching, interactions, feedback, given outputs, etc. On the use and user level, questions of application to given problems and tasks are raised. The assessment criteria are market and fitness-of-use. And finally, on the social level, issues of impact on the environment(s) occur. The evaluation criteria include effects on research, productivity, and decision-making.

Van House (1995) describes user needs assessment and evaluation in the Berkeley NSF/NASA/ARPA Digital Library Initiative project. Five levels of analysis are defined: environment (e.g., organizational, political, social, professional); task-general (tasks of an organization); task-specific (tasks of an individual user or group of users); information acts; and digital library use (actual interaction with the digital library).

Bishop (1995, cited by Harter & Hert, 1997, p.48) reports on the work of the User Research Working Group for the entire federal Digital Library Initiative, which has been developing methods both for investigating users and for evaluating usage. Evaluation layers developed by the group are adequacy of the collection, functionality, interface and usability, search and retrieval performance and behavior, effect on work of users, fundamental changes in processes, and public policy implications.

These frameworks suggest that different types of evaluation are necessary. Is there a way to synthesize these frameworks and paradigms in IR evaluation and develop measures and metrics that can be applied to individual system evaluations and used across systems to compare systems features?

4.3 Synthesizing frameworks and paradigms

Generally speaking, the overall purpose of IR evaluation is to improve the probability that any given IR system will be adopted and used. Rogers (1995) has identified five attributes of innovations that have been correlated to the adoption of innovations. The first attribute, relative advantage, is the degree to which a new innovation supersedes current practices. It can be operationalized, or measured in terms of, variables such as speed, economic gain, added convenience, and social prestige. The second attribute, compatibility, is the degree to which an innovation is perceived to be consistent with adopters' values, experiences, and future needs. Compatibility with the social structure, individual and associated group beliefs, organizational or social climate, and individual and group goals have been shown to influence adoption of innovations. The third attribute is complexity. Complexity refers to the difficulty of learning to use and understand a new system or technology. It can be measured by the number of new skills and/or knowledge needed in order to use and benefit

from an innovation. The fourth attribute, trialability, refers to the ease of experimenting with an innovation on a limited basis, and includes the level of effort needed and risk involved in observing and participating in small scale demonstrations of the system, and the costs involved in reversing, or stopping to use the system. The fifth attribute is observability, the degree to which the results of the innovation are observable. These attributes have been studied for a variety of innovations and situations (e.g., Tornatzky & Klein, 1982; Rogers, 1995). In different situations and contexts, some attributes appear to be more important than others are. However, in general, all attributes appear to be influential.

The evaluation paradigms and frameworks described earlier address aspects of the five attributes of innovation to varying degrees. For example, recall and precision used in the systems-oriented paradigm measure one aspect of relative advantage. Socio-technical design and evaluation focuses on aspects of compatibility. Usability engineering, cognitive systems engineering, Belkin' s user level measure aspects of complexity. Belkin' s access level is related to complexity, trialability and observability. We propose that the attributes can be used as the foundation for evaluation. The attributes represent values that are important for the adoption and use of innovations from the perspective of the adopters, and thus provide face and theoretical validity to the evaluation framework that is lacking in frameworks that divide evaluation into levels which represent distinctions based on designers' perspectives.

Table 1 illustrates a possible evaluation framework based on the attributes. Criteria and measures for each attribute are proposed**.** In the proposed framework the first attribute, relative advantage, has the criterion of system relevance, topical relevance, speed and economic gain. These criteria can be measured through recall and precision, the scope of the source contents, system response time, and a cost benefit analysis. The second attribute, compatibility, can be operationalized by three criteria: motivational relevance, organizational relevance, and social relevance. These criteria, in turn, can be measured by how well the IR system meets users' expectations, (associated) organizations' expectations, and society' s expectations. Social relevance can also be measured by investigating the system' s compatibility with public policy.

A third attribute of innovation adoption, complexity, has the criteria of usability, cognitive relevance, and situational relevance. Measures associated with these criteria include task completion time, error rate, error correction time, task completion rate, user satisfaction, and user satisfaction in problem or work contexts. The fourth attribute, trialability, has the criterion of ease of experimentation that can be measured by availability, training time, and other start-up costs calculations. The fifth attribute, observability, has the criterion, degree

of demonstration that can be measured by the cost of observation.

Table 1. Proposed evaluation framework

Attribute	**Criteria**	**Measures**
Relative advantage	System relevance Topical relevance Speed Economic gain	Recall & precision Source contents (type & coverage) System response time Cost benefit analysis
Compatibility	Motivational relevance Organizational relevance Social relevance	Meets users' expectations Meets organizations' expectations Meets society' s expectations, Compatibility with public policy
Complexity	Usability Cognitive relevance Situational relevance	Task completion time, error rate, error correction time Task completion rate, user satisfaction User satisfaction in problem or work contexts
Trialability	Ease of experimentation	Availability, training time, & other start-up costs
Observability	Degree of demonstration	Cost of observation

A variety of techniques can be used to calculate the measures, including IR laboratory experiments to measure recall, precision, and system response time, and usability engineering lab experiments to measure task completion time, error rate, error correction time, and training time. In addition, a combination of interviews, observations, and surveys can be used to determine user satisfaction, availability, cost of observation, and how well the IR system meets expectations. Operation research methods can be used to determine the cost/benefit ratio for the system. We propose that no one technique is appropriate for evaluation but a combination is required to do a comprehensive evaluation.

We can not claim that Table 1 is complete, i.e., that all criteria, measures, and techniques have been identified. However, we believe that the attributes provide a valid, theoretical framework on which to begin to synthesize the IR evaluation paradigms and other evaluation frameworks that have emerged in IR and other fields. Furthermore, we do not claim that all criteria and measures are necessary for all contexts. Designers and developers of IR systems

must decide which attributes and criteria are most important for their context. For example, compatibility with public policy may not be an important measure for an intranet-based IR system. We propose that the framework begins to identify the options available. Additional research is required to ascertain the usefulness of this approach.

5. Summary

Systems-oriented research and user studies have made many contributions to the field of information retrieval, e.g., their results have led to improved IR systems and services. Within the systems-oriented community, TREC, the Japanese BMIR-J2 IR system, and C-TREC are endeavors that extend IR evaluation to include new very large document collections and different types of retrieval tasks. Within the user studies community, conferences and professional groups have emerged to assist researchers in sharing knowledge about research methods and results.

Synthesizing these approaches promises to provide a more comprehensive approach to IR evaluation. In this paper we propose a framework that synthesizes the approaches. The framework is based on attributes that have been shown to influence adoption of innovations. Evaluation criteria and measures are derived from these attributes. This approach may provide a framework that helps to design, or formulate, evaluation research and synthesize evaluation results across different systems, situations and contexts. Further work is necessary to demonstrate the utility of the framework.

Acknowledgments

We would like to thank Bob Losee for his comments on a draft of this paper.

References

Algon, J. (1996). Classification of tasks, steps, and information-related behaviors of individuals on project teams. In P. Vakkari, R. Savolainen, & B. Dervin (Eds.). *Information Seeking in Context* (pp. 205-221). London: Taylor Graham.

Beer, S. (1974.) *Designing Freedom*. Great Britain: Garden City Press, Ltd.

Belkin, N. (1994). Digital libraries: interaction and evaluation. Talk given in the Information Retrieval as Interaction Workshop. Academia Sinica, Taipei, August 30-31, 1994.

Bishop, A. P. (1995). Working towards an understanding of digital library use. D-Lib Magazine. 1995 October. http://www.dlib.org/dlib/october95/10bishop.html

Borgman, C., Hirsh, S., & Gallagher, A. (1995). Children' s' searching behavior in browsing and keyword online catalogs: The Science Library Catalog Project. *Journal of American Society for Information Science, 46* (9), 663-378.

Borlund, P. & Ingwersen, P. (1997). The development of a method for the evaluation of interactive information retrieval systems. *Journal of Documentation, 53* (3), 173-94.

Chatman, E.A. (1996). The impoverished life-world of outsiders. *Journal of American Society for Information Science, 4 7*(3), 193-206.

Cheng, Y. & Shaw, D. (1999). Information seeking and finding. *Bulletin of the American Society for Information Science, 25* (3), 10-11.

Ellis, D. (1996). The dilemma of measurement in information retrieval research. *Journal of the American Society for Information Science, 47* (1), 23-36.

Gluck, M. (1996). Exploring the relationship between user satisfaction and relevance in information systems. *Information processing & Management, 32* (1), 89-104.

Harter, S. P. (1996). Variations in relevance assessments and the measurement of retrieval effectiveness. *Journal of the American Society for Information Science, 47* (1), 37-49.

Harter, S. & Hert, C. (1997). Evaluation of information retrieval systems: Approaches, issues, and methods. M. Williams (Ed.), *Annual Review of Information Science and Technology, Vol. 32 (pp. 3-94).* Medford, N. J: Information Today.

Iivonen, M. (1995). Consistency in the selection of search concepts and search terms. *Information Processing & Management, 31* (2), 173-183.

Iivonen, M / & Sonnenwald, D. (1998). From translation to navigation of different discourses: A model of search term selection during the pre-online stage of the search process. J*ournal of the American Society for Information Science, 49* (4), 312-326.

Ingwersen, P. (1997). Cognitive perspectives of information retrieval interaction: Elements of a cognitive information retrieval theory. *Journal of Documentation, 52* (1), 3-30.

Kuhlthau, C. (1996). The influence of uncertainty on the information seeking behavior of a securities analyst. In P. Vakkari, R. Savolainen, & B. Dervin (Eds.). *Information Seeking in Context* (pp. 268-274). London: Taylor Graham.

Kuhlthau, C. (1993). *Seeking Meaning.* Greenwich, CT: Ablex Publishing Co.

Lancaster, F. W. et al. (1996). Evaluation of interactive knowledge-based systems: overview and design for empirical testing. *Journal of the American Society for Information Science, 47* (1), 57-69.

Losee, R. M. (1996). Evaluating retrieval performance given database and query characteristics: analytic determination of performance surfaces. *Journal of the American Society for Information* Science, 47 (1), 95-105.

Mumford, E. (1995). *Effective Systems Design and Requirements Analysis: The ETHICS Approach.* London: Macmillan.

Pejtersen, A. (1989). A library system for information retrieval based on a cognitive task analysis and supported by an icon-based interface. In N. Belkin & C. van Rijsbergen (Eds.), *Proceedings of the 12th Annual International ACM SIGIR Conference* (pp.40-47). New York: ACM.

Rasmussen, J., Pejtersen, A.M., & Goodstein, L.P. (1994). *Cognitive Systems Engineering.* New York: Wiley.

Robertson, S.E. & Beaulieu, M. (1997). Research and evaluation in information retrieval. *Journal of Documentation , 53* (1), 51-57.

Rogers, E. (1995). *Diffusion of Innovations*. New York: The Free Press.

Salton, G. & McGill, M.J. (1983). *Introduction to Modern Information Retrieval*. New York: McGraw-Hill.

Saracevic, T. (1997). User Lost. *ACM SIGIR Conference Proceedings.* (ftp://scils.rutgers.edu/pub/tefko/relevance.doc)

Saracevic, T. (1996). RELEVANCE reconsidered 1996. *Proceedings of the 2nd International Conference on the Conceptions of Library and Information Science (CoLIS2).* Copenhagen,

Denmark, 14-17, Cot.1996.

Saracevic, T. (1995). Evaluation of evaluation in information retrieval. *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (*pp138-146.) New York: ACM.

Saracevic, T. (1975). RELEVANCE: A review of and a framework for the thinking on the notion in Information Science. J*ournal of the American Society for Information Science* (Nov/Dec, 1975), 321-343.

Saracevic, T., & Kantor, P. (1988a). A study of information seeking and retrieving. II. Users, questions and effectiveness. *Journal of American Society for Information Science, 39,* 177-196.

Saracevic, T., & Kantor, P. (1988b). A study of information seeking and retrieving. III. Searchers, searches, and overlap. *Journal of American Society for Information Science, 39*, 197-216.

Schamber, L. (1994). Relevance and information behavior. In: M.E. Williams (Ed.), *Annual Review of Information Science and Technology, vol. 29,* (pp. 3-48.) Medford, N. J: Information Today.

Schamber, L., Eisenberg, M., & Nilan, M. (1990). A re-examination of relevance: toward a dynamic, situational definition. *Information Processing & Management, 26* (6), 755-756.

Solomon, P. (1993). Children' s information retrieval behavior: A case analysis of an OPAC. *Journal of American Society for Information Science, 44* (5), 245.

Solomon, P. (1997). Discovery of information behavior in sense making. *Journal of American Society for Information Science, 48* (12), 1097-1138.

Sonnenwald, D.H. (1999). Evolving perspectives of human information behavior: Contexts, situations, social networks and information horizons. In T. Wilson (Ed*.), Information Seeking in Context* (vol.2.) New York: MacGraw Hill.

Sonnenwald, D.H. (1996). Communication roles that support collaboration during the design process. *Design Studies, 17,* 277-301.

Sonnenwald, D.H., & Lievrouw, L.A. (1997). Collaboration during the design process: A case study of communication roles and project performance. In P. Vakkari, R. Savolainen & B. Dervin (Eds.), *Information Seeking in Context* (pp. 179-204). London: Taylor Graham.

Sonnenwald, D.H., Marchionini, G., Wildemuth, B. M., Dempsey, B.L., Viles, C.R., Tibbo, H.R., Smith, J.S. (1999). Collaboration services in a participatory digital library: An emerging design. *Third International Conference on Conceptions of Library and Information Science*. (to appear.)

Sonnenwald, D.H., & Pejtersen, A.M. (1994). Towards a framework to support information needs in design: A concurrent engineering example. In H. Albrechtsen & S. Oernager (Eds.). *Knowledge Organization and Management* (pp. 161-172). Frankfurt/Main: Indeks Verlag, 1994.

Sonnenwald, D.H., & Pierce, L. (1999). Information behavior in dynamic group work contexts: Interwoven situational awareness, dense social networks, and contested collaboration in command and control. *Information Processing & Management.* (in press).

Sparck Jones, K. (1981). The Cranfield tests. In K. Sparck Jones (Ed.), *Information Retrieval Experiment* (pp. 256-284). London: Butterworths.

Spink, A., Wilson, T., Ellis, D., & Ford, N. (1998). Modeling users' successive searching: A National Science Foundation/ British Library Study. *D-Lib Magazine* (April).

Su, L. (1992). Evaluation measures for interactive information retrieval. *Information Processing & Management, 28* (4), 503.

Tague-Sutcliffe, J. (1996a). Information retrieval experimentation. in A. Kent (Ed.) *Encyclopedia of library and information science, v.57 Supplement 20* (pp.195-209). New York: Marcel Dekker.

Tague-Sutcliffe, J. (1996b). Some perspectives on the evaluation of information retrieval systems. *Journal of American Society for Information Science, 47* (1), 1-3.

Tornatzky, L.G., & Klein, K.J. (1982). Innovation characteristics and innovation adoption implementation: A meta-analysis of findings. *IEEE Transactions on Engineering*

Management, EM-29, 28-45.

TREC (1999). Text Retrieval Conference Home Page. http://trec.nist.gov/

Vakkari, P., Savolainen, R., & Dervin., B. (Eds.). (1997). *Information Seeking in Context*. London: Taylor Graham.

Van House, N. (1995). User needs assessment and evaluation for the UC Berkeley Electronic Environmental Library Project: A preliminary report. In F. M. Shipman, R. K Furuta, & D. M. Levy (Eds.), *Proceedings of Digital Libraries '95: The 2nd annual conference on the Theory and Practice of Digital Libraries* (pp.71-76). Texas: A&M University. http://csdl.tamu.edu/DL95/papers/vanhouse/vanhouse.html

Voorhees, E. (1998). Variations in relevance judgments and the measurement of retrieval effectiveness. *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (*pp. 315-323). New York: ACM.

Whiteside, J., Bennett, J., & Holtzblatt, K. (1988). Usability engineering: Our experiences and evolution. In M. Helander (Ed.), *Handbook of Human-Computer Interaction* (pp. 791-817). New York: Elsevier Science Publishers.

Wilson, T. (1997). Information behavior: An inter-disciplinary perspective. In P. Vakkari, R. Savolainen & B. Dervin (Eds.), *Information Seeking in Context* (pp. 39-52). London: Taylor Graham.

Wilson, T. (Ed.), (1999). *Information Seeking in Context* (vol.2.) NY: MacGraw Hill.

Wu, M.M. (1998). Relevance judgment: The notions of logical relevance and pertinence. *Journal of Library & Information Science, 24* (2), 44-64. (In Chinese)

Wu, M.M. (1998). *An Evaluation Test-bed for Chinese Information Retrieval.* Technical Report. Funded by Industrial Information Institution (III).