Relevance and Retrieval Evaluation: Perspectives from Medicine

William Hersh
Biomedical Information Communication Center, Oregon Health Sciences University, 3181 SW Sam Jackson Park Road, Portand, OR 97201

The traditional notion of topical relevance has allowed much useful work to be done in the evaluation of retrieval systems, but has limitations for complete assessment of retrieval systems. While topical relevance can be effective in evaluating various indexing and retrieval approaches, it is ineffective for measuring the impact that systems have on users. An alternative is to use a more situational definition of relevance, which takes account of the impact of the system on the user. Both types of relevance are examined from the standpoint of the medical domain, concluding that each have their appropriate use. But in medicine there is increasing emphasis on outcomes-oriented research which, when applied to information science, requires that the impact of an information system on the activities which prompt its use be assessed. An iterative model of retrieval evaluation is proposed, starting first with the use of topical relevance to insure documents on the subject can be retrieved. This is followed by the use of situational relevance to show the user can interact positively with the system. The final step is to study how the system impacts the user in the purpose for which the system was consulted, which can be done by methods such as protocol analysis and simulation. These diverse types of studies are necessary to increase our understanding of the nature of retrieval systems.

Introduction

The debate over the meaning of relevance has significant implications for research in information retrieval. In most retrieval evaluation studies, relevance is assumed to mean topicality. A system or a user is measured by how many documents that topically relate to the query are retrieved for one or more queries. This approach to evaluation was pioneered by Cleverdon and Keen (1966). By creating test collections with fixed queries, documents, and relevance judgments, many different approaches to indexing and retrieval can be tested, often without the need for expensive and time-consuming experiments with users. The major modern proponent of this approach is Salton (1992). But as users, designers, and evaluators of systems, we know that there is more to consider in a retrieval system than just the number of relevant and nonrelevant documents retrieved. Systems must also be easy to use, helpful in constructing queries, and able to steer users to the types of documents they need. They also have to contain appropriate breadth and depth of information for users to obtain value. This has led to a search for broader definitions of relevance, many of which are considered in the review paper by Saracevic (1975), who noted the paradox of the centrality of the concept of relevance versus our inability to define it.

An emerging alternative view of relevance that has existed and is generating renewed interest is that of situational relevance. This notion of relevance, most recently advocated by Schamber, Eisenberg, and Nilan (1990), abandons the focus on the information retrieval system, looking instead at the impact of that system upon the information needs of the individual user. The number of relevant articles retrieved is still among the parameters to be measured, but a relevant article is defined not as merely on the topic of the user's search, but rather as providing information that can be used. Harter (1992) has used the term "psychological relevance" to describe this, stating that relevant documents must either give the user new knowledge, correct old knowledge, or update existing knowledge.

There is much experimental evidence for this approach. To begin with, judgments of topical relevance themselves can be inconsistent, especially when applied by nonusers (Swanson, 1988). In addition, the traditional method of judging relevance into fixed categories may be flawed, with numerical categories providing a more consistent approach (Eisenberg, 1988). More importantly, others have shown that there are more factors that go into a document being relevant than just topicality. Cuadra and Katter (1967) showed that relevance for a user was influenced not only by subject matter, but also by the needs, prejudices, and knowledge of the user. Cooper (1973), arguing that topical relevance could not take into account the utility of a user's interaction with a system, defined a measure of utility based upon "whatever the user finds to be of value about the system output, whatever its usefulness, its entertainment, or aesthetic value, or anything else."

But is situational relevance sufficient for the proper evaluation of retrieval systems? In this article, we look at the retrieval evaluation problem from the standpoint of medical research, using known flaws in the medical literature to demonstrate that providing the user with topical information *and* in a manner that changes their knowledge state is not enough. In medicine, there is increasing emphasis on *outcomes-oriented* research (Littenberg, 1992) where, for example, a therapy must be shown not only to improve the value of a blood test or reduce symptoms, but also to lead to long-term benefit for the patient, especially when compared with other therapies or none at all. We apply this approach to information retrieval systems, concluding that systems should ultimately be judged by how well they help users in their task for which they consult the system, whether it is to make correct decisions or lead to some improved state in what they do with the information. An outcomes-based approach provides a different perspective on the topical versus situational relevance debate, indicating that neither is adequate for a complete assessment of retrieval systems. Each type of relevance has its shortcomings, and since relevance is central to evaluation of retrieval systems, use of one type alone in a retrieval evaluation will lead to inadequate results.

The central thesis of this article is that deciding on a definition of relevance depends on the context in which it is being applied. From the standpoint of evaluation, the type of relevance assumed need not be an either-or choice. Topical relevance is useful for certain types of evaluation, and situational relevance is useful for others. Ultimately, however, we must search for measures of relevance and retrieval that enable us to demonstrate an improved outcome resulting from the use of retrieval systems.

Topical and Situational Relevance in the Context of Medical Literature

Before considering relevance in medical information retrieval systems, one must understand the myriad of information problems in medicine. Despite societal perceptions to the contrary, physicians often make decisions based on inaccurate and/or incomplete knowledge. Unfortunately, the extent to which this impacts clinical care has never been measured (and, in fact, would be very hard to do). It has been shown, for example, that infectious disease experts disagree with 25–50% of antibiotic choices made by primary care physicians (Bernstein, Barriere, & Conte, 1982; Kunin, Tupasi, T., & Craig, 1973). The impact of this on actual patient outcomes is unknown and may be impossible to measure. Nonetheless, numerous studies have documented deficiencies in physician information. The most frequently cited of these is by Covell, Uman, and Manning (1985), who showed that physicians drastically underestimate how often they have unmet information needs. Covell studied a group of generalist and specialist internists who estimated having unmet needs once per month. When actually observed, however, they turned out to have two unanswered questions for every three patients. In a different study, Williamson et al. (1989) surveyed practicing physicians for their knowledge on important clinical advances that had recently been described in a major medical journal, finding that as many as 50% of physicians were unaware of at least one of the significant recent advances identified.

The above studies have been used to advocate increased emphasis on improving physician access to information, with some of this advocacy devoted to increasing development and use of end-user retrieval systems. Physicians are using retrieval systems in increasing numbers and are no doubt retrieving relevant articles, whether they be relevant by topical or situational considerations. While these systems are delivering information to these users, and changing their knowledge state, the utility of this information is uncertain. But before we can begin to assess the impact of retrieval systems in medicine, we must consider the validity of medical information itself. The last half-century has seen an explosion of medical knowledge based on scientific studies. Yet only in the last couple decades has adequate attention been paid to the scientific quality of those studies.

Scientific Validity

The scientific validity of medical research is important to the discussion of relevance, since just because an article meets the criteria for topical or situational relevance, it does not mean that the article contains conclusions warranted by the data. Fletcher and Fletcher (1979) found that poor methodology plagues a considerable portion of medical research, leading to questions about its validity. DerSimonian et al. (1982) discovered that reporting of methods in the medical literature is insufficient, indicating that readers of articles may not be able to judge whether the data support the conclusions. Glantz (1980) showed that over a quarter of studies published in an esteemed cardiology journal misused the t test in their analyses. Bailar (1986) cast a very critical tone on medical research in general, noting that few papers report their inadequacies, probably over the authors' fears of rejection. Also of serious concern is outright research fraud. Science is not immune to fallabilities in human character, as demonstrated by Kochan and Budd (1992), who recently investigated citations of the work of John Darsee, a Harvard researcher found to have committed fraudulent research yet whose other work is still cited positively.

The high incidence of flawed research shows that both topical and situational relevance are not enough to evaluate the effect of retrieval systems on clinical end-users. An article may be on the topic, and may even change the knowledge state of the user, but if its conclusions are not justified, then the information flow is ineffective. There is increasing emphasis in medical education on critical appraisal of medical research (Evidence-Based Medicine Working Group, 1992), a process in which most physicians are unskilled. This fact may lead one on to consider a position diametrically opposed to the situational relevance model, since it may be possible that a user is distinctly

unqualified to judge relevance on the articles he or she has retrieved.

Other Issues in Scientific Literature

Another problem in information being misconveyed is a poor abstract. This is exacerbated by the increasing availability of bibliographic retrieval systems outside of libraries, where the user may not have the time or motivation to seek the full article. It has been shown that abstracts may inadequately summarize the article or be outright misleading (Haynes et al., 1990b). This presents a problem for both the topical and situational views of relevance, since a reference seemingly relevant by the abstract may not be so, or vice versa, altering a user's decision to seek the article. Thus poor writing could affect the performance of a retrieval system through no fault of the indexers or system designers.

Even if an article's conclusions are warranted, an additional dilemma is that of changing knowledge. In all fields, science is an iterative, progressive process, with new knowledge building on previous foundations. Retrieval systems may only obtain parts of the picture, obscuring the evolutionary process. One need only look at the evolution of medical thinking about the role of serum cholesterol in heart disease over the years (Littenberg, 1992). When the link between cholesterol and heart disease was first discovered, the general consensus of the medical community was that altering the cholesterol level would have no impact on development of heart disease. Eventually, clinical studies were done showing that lowering the cholesterol level did indeed reduce the risk of heart disease. However, it has become clear in recent years that not everyone may benefit from testing for and lowering cholesterol levels, and in fact it is not cost-effective to screen and treat the entire population. The significance of this example to retrieval systems is that a user may retrieve many relevant articles, but they may not reflect the entire picture of a scientific area in evolution.

Limitations of Current Evaluation Methodology

Recall and Precision

The limitations of both topical and situational relevance in the context of scientific validity and evolution discussed above imply serious limitations in the utility of the most commonly used measures for retrieval system evaluation, recall and precision. As information scientists, we lament the 20–50% recall that is obtained in most studies. But what is the significance of a difference in recall and precision? Haynes et al. (1990a) showed that expert MEDLINE searchers, whether librarians or clinicians, had a level of recall twice that of clinicians who were novice searchers (48–49% vs. 27%). Relevance was measured by topicality as judged by physicians on a medical school faculty. One surprising finding in this study was the almost total lack of overlap in the sets of relevant articles retrieved. Thus, even though the novices found fewer relevant articles, they found some articles that expert searchers did not. Unfortunately this study did not examine the impact of the searches on the users' state of knowledge, though that would have been very hard to do.

Another interesting aspect to Haynes' study was the general enthusiasm and satisfaction for searching by the novices, despite only obtaining half the recall of the experts. Did the searchers obtain useful information from the articles they retrieved? There is, after all, a great deal of redundancy in the medical literature and it is possible that the articles they retrieved answered their questions. Or it is possible that by reference tracing or some other mechanism they were able to discover other relevant articles. Of course it is also possible that they were blissfully ignorant, falsely assuming they were retrieving the bulk of the literature on their topics.

This study highlights the limitations of using recall and precision alone to evaluate retrieval systems. (Haynes did not do this, but many evaluators do.) These measures are important, but they must be placed in perspective, and particular attention must be paid to their meaning. Again, medical research offers some perspective here, particularly on the notion of comparison of differences. When discussing differences it is important to distinguish between clinical and statistical significance. For example, when comparing the efficacy of a drug for treating diabetes, one can measure the improvement in blood sugar level. Let us say that one drug shows a lowering of the blood sugar by 10 mg/dl, a very modest amount. With a large enough sample size, this can be shown to be a statistically significant difference. But is this a clinically significant difference? The magnitude of the difference is small, and probably does not contribute to a patient's well-being or long-term outcome. This same perspective must be maintained when interpreting differences in recall and precision from retrieval studies. We all know that there is a meaningful difference between a system achieving 80% recall versus one obtaining 20% recall, but do not know the practical significance of, say, a difference between 50% and 60% recall. Or a 10% difference in precision at a fixed point of recall. We may be able to produce a sample size with the power to show statistical significance, but such a difference says nothing about whether the user obtained articles that enable him or her to get the desired information.

Experimental Setting

Another variable in retrieval evaluation that must be reconsidered in light of the previous discussion is the experimental setting. A large line of evaluation is based upon research where data measurements consist of performing experiments in noninteractive, batch mode (Salton, 1992). To evaluate a system with this approach, a database, a set of queries, and relevance judgments between queries and documents are used. While this methodology has utility in the early phases of system design, it is unrealistic to make claims that a certain

approach to indexing or retrieval can be judged based on these types of experiments alone. For example, Salton et al. (1990) dismiss the use of natural language processing techniques, based entirely on experiments of this type. This ignores the benefit that these techniques may have in helping a user formulate a query, or providing a more fertile substrate for relevance feedback. Batch mode studies can be useful for a preliminary evaluation of a particular indexing or retrieval algorithm, but studies with computer–human interfaces and real users must be performed before proclamations can be made as to the fitness of different algorithms.

Concerns about experimental setting also must take into account the nature of queries and databases. Going back to medical experiments, we must remember that the population as well as the setting in which the population is studied are important variables. For example, many presentations of diseases have distinct ethnic and gender differences. One cannot make reliable conclusions about the effect of a disease on one group by extrapolating from the other. Furthermore, patients who are studied in one setting, such as a large referral center at a medical school, are likely to be different than those observed in another, such as a community clinic. Patients sent to referral centers are likely to have more complex presentations and advanced progression of a disease. This same perspective must be acknowledged in retrieval studies. For example, queries that express the interests of researchers are unlikely to help in evaluating systems designed for clinicians. Thus if we study a group of end-users from one profession or even one group within a profession, our results are likely to be only valid for that group. To make more general conclusions, we must study many groups and look for similarities.

A Framework for Future Research

It is clear from the problems with medical information in general that both topical and situational relevance alone are inadequate for comprehensive evaluation of medical (and probably all) information systems. Topical relevance ignores the change in state of the user's knowledge when he or she interacts with the system. Meanwhile, situational relevance focuses only on the user's interaction with the system as a whole, making assessment of individual components (such as indexing or retrieval algorithm) more difficult. But even though both types of relevance are inadequate, this does not mean they are useless in retrieval evaluation. In fact, as will be argued below, both are important and lead to valid research. In this section we provide a framework for retrieval research, developing an iterative approach that starts with the use of topical relevance, proceeds next with assessment of situational relevance, and ultimately uses outcomes-based methods to assess the impact of the system upon the user for the task which he or she has sought to use the system.

Role of Topical Relevance

For assessing different approaches to indexing and retrieval, topical relevance is very useful for insuring that users can get access to the subject material they need to with the best recall and precision. Studies using a batch mode for evaluation can be excellent preliminary studies to isolating certain aspects of indexing or retrieval. For example, if a new approach to indexing using natural language processing were developed, it would be important to know whether the indexing process yields representations that can result in adequate retrieval. The mistake is too often made to end evaluation and draw conclusions at this level.

Role of Situational Relevance

To assess the impact of retrieval systems on actual users, however, situational relevance must be used. Recent studies by Katzer and Snyder (1990) and Osheroff and Bankowitz (1993) incorporate the new thinking on situational relevance in better assessing retrieval system performance. This type of relevance is also very useful for evaluation of innovative interfaces that attempt to model the user or provide relevance feedback. It may even be important to combine measures of topical and situational relevance. In an evaluation study we did of the SAPHIRE system, questionnaire responses from users were as important as recall and precision, and additional effort was made to correlate the two (Hersh & Hickam, 1992).

Or course, as we have seen, even situational relevance is not enough, since users may have their information state worsened by the retrieval of erroneous or misleading documents. Furthermore, a system with exemplary indexing and retrieval capabilities may be shunned by a user who needs a different type of document or database. We may have a perfectly good retrieval system that happens to fare poorly in an evaluation possibly due to a skewed set of users. Ultimately, each component as well as the entire flow of information must be studied in order to assess the impact of computer-based retrieval systems. This leads us to develop the next step, an outcomes-oriented framework for evaluation.

Toward Outcomes-Oriented Evaluation

An outcomes-based approach to retrieval evaluation presents many challenges. It requires that we go beyond the closed world of the system, or of the system and the user interacting with it. We must find meaningful measures of assessing how valuable a system can be to the user, whether he or she is a professional making decisions or a researcher looking for new information on a potential research idea. Again we can borrow from research methodology used in medicine. These methods, while not unique to medicine, have been used successfully by medical education researchers to evaluate physician decision making, looking at how decision making can itself be assessed and how interventions to change it can be

measured. Two general approaches have been developed, each of which could be used in evaluating the impact of retrieval systems.

The first of these is protocol analysis (Ericsson & Simon, 1984), which has been used in assessing medical problem-solving (Elstein, Shulman, & Sprafka, 1978a; Kassirer & Gorry, 1978). In this type of study, a user is monitored while performing a given task. Actions are recorded, and the user may be prompted to "think aloud." In the case of retrieval systems, the entire retrieval process would be recorded, from the original statement of information need and first interaction with the searching system, to the final assessment of all the information retrieved and how much impact it has had on the user's work. In a clinical practice setting, for example, an approach would be to follow a clinician in his or her practice, closely monitoring interactions with the retrieval system and assessing its impact upon decisionmaking. The session would be reviewed by the experimenter to determine such factors as whether the system provided answers to questions that had an impact in patient care, whether someone more expert in the use of the system could have obtained better answers to questions, and, in the case of systems with more than one database, whether the most appropriate ones were utilized.

The second type of approach that could be used is simulation (Elstein et al., 1978b; Schwartz & Griffin, 1986), using instruments such as questionnaires and patient management problems. The latter are tests that present a patient-care scenario which are graded by the user's selection of diagnostic tests (from the physical examination to expensive or invasive tests) and treatments (from simple and safe medications to complicated surgeries). Each test-taker starts at the same point, but based upon the pathways chosen (such as deciding to do an invasive test or treatment) different outcomes may ensue. The score on this type of examination reflects the appropriateness of selection of diagnostic tests and therapeutic interventions. Similar to protocol analyses described above, patient management simulations have been used successfully to assess physician competence. These simulations could be used to evaluate retrieval systems by giving the same problems to groups of users who have access to different types of retrieval systems (or even none at all) and looking for differences between groups. Using simulation for testing is hardly unique to medicine, and similar approaches could be developed in other domains. Of course the methods of protocol analysis and simulation are not mutually exclusive and most of the studies cited above combined both.

Are there any drawbacks to outcomes-oriented evaluation? It is certainly more expensive and time-consuming than simple batch runs in a laboratory or replication of end-user searches by experts. And it would definitely be impractical to have to do these kinds of studies to test every new approach to indexing and retrieval. However, once an approach to retrieval design has shown its mettle in batch studies using topical relevance and initial user studies using situational relevance, the next step of an outcomes-based approach must be taken to insure that the system truly has value to a user.

Conclusion

We have explored the limitations of topical and situational relevance with examples from the medical domain that lead us to develop a new model for retrieval evaluation based on an iterative, outcomes-oriented approach, which allows assessment of retrieval systems in the context of the task which the user has chosen to use the system. This approach takes advantage of the beneficial aspects of each type of relevance, but uses them in combination to nullify their individual weaknesses. Early evaluations of systems should focus on topical relevance. A system must be able to retrieve topically relevant documents before anything else can be assessed. But no evaluation of a system can be complete without measuring its impact on users. The next step is to measure the impact of systems upon users, with particular emphasis paid on steering users to documents that change their knowledge state. These types of evaluation are best measured by situational relevance. Ultimately, however, we must measure the outcome of the whole system–user interaction, which leads beyond the notions of relevance entirely. As we move toward an outcome-based assessment of retrieval systems for each group of users, we must start to integrate the whole picture and determine what aspects of systems are important to all users versus just helpful for certain groups. This will ultimately lead us to understand the nature and role of retrieval systems.

Acknowledgments

The author acknowledges the comments and suggestions on early drafts from David H. Hickam, MD, MPH, R. Brian Haynes, MD, PhD, and David A. Evans, PhD.

References

Bailar, J. (1986). Science, statistics, and deception. *Annals of Internal Medicine, 104,* 259–260.

Bernstein, L., Barriere, S., & Conte, J. (1982). Utilization of antibiotics: Analysis of appropriateness of use. *Annals of Emergency Medicine, 11,* 21–24.

Cleverdon, C., & Keen, E. (1966). *Factors determining the performance of indexing systems* N. Cranfield, UK: Aslib Cranfield Research Project.

Cooper, W. (1973). On selecting a measure of retrieval effectiveness, part 1: The subjective philosophy of evaluation. *Journal of the American Society for Information Science, 24,* 87–100.

Covell, D., Uman, G., & Manning, P. (1985). Information needs in office practice: Are they being met? *Annals of Internal Medicine, 103,* 596–599.

Cuadra, C., & Katter, R. (1967). Opening the black box of "relevance." *Journal of Documentation, 23,* 291–303.

DerSimonian, R., Charette, L., McPeek, B., & Mosteller, F. (1982). Reporting on methods in clinical trials. *New England Journal of Medicine, 306,* 1332–1337.

Eisenberg, M. (1988). Measuring relevance judgments. *Information Processing and Management, 24,* 373–389.

Elstein, A., Shulman, L., & Sprafka, S. (1978a). Effects of hypothesis generation and thinking aloud. In *Medical problem solving: An analysis of clinical reasoning* (pp. 228–251). Cambridge, MA: Harvard University Press.

Elstein, A., Shulman, L., & Sprafka, S. (1978b). High-fidelity simulation: Research methods. In *Medical problem solving: An analysis of clinical reasoning* (pp. 46–63). Cambridge, MA: Harvard University Press.

Ericsson, K., & Simon, H. (1984). *Protocol analysis: Verbal reports as data.* Cambridge, MA: MIT Press.

Evidence-Based Medicine Working Group. (1992). Evidence-based medicine: A new approach to teaching the practice of medicine. *Journal of the American Medical Association, 268,* 2420–2425.

Fletcher, R., & Fletcher, S. (1979). Clinical research in general medical journals: A 30–year perspective. *New England Journal of Medicine, 301,* 180–183.

Glantz, S. (1980). How to detect, correct, and prevent errors in the medical literature. *Circulation, 61,* 1–7.

Harter, S. (1992). Psychological relevance and information science. *Journal of the American Society for Information Science, 43,* 602–615.

Haynes, R., McKibbon, K., Walker, C., Ryan, N., Fitzgerald, D., & Ramsden, M. (1990a). Online access to MEDLINE in clinical settings. *Annals of Internal Medicine, 112,* 78–84.

Haynes, R., Mulrow, C., Huth, E., Altman, D., & Gardner, M. (1990b). More informative abstracts revisited. *Annals of Internal Medicine, 113,* 69–76.

Hersh, W., & Hickam, D. (1992). A comparison of two methods for indexing and retrieval from a full-text medical database. In *Proceedings of the 55th Annual Meeting of the American Society for Information Science,* pp. 221–230.

Kassirer, J., & Gorry, G. (1978). Clinical problem solving: A behavioral analysis. *Annals of Internal Medicine, 89,* 245–255.

Katzer, J., & Snyder, H. (1990). Toward a more realistic assessment of information retrieval performance. In *Proceedings of the 53rd Annual Meeting of the American Society for Information Science,* pp. 80–85.

Kochan, C., & Budd, J. (1992). The persistence of fraud in the literature: The Darsee case. *Journal of the American Society for Information Science, 43,* 488–493.

Kunin, C., Tupasi, T., & Craig, W. (1973). Use of antibiotics—a brief exposition of the problem and some tentative solutions. *Annals of Internal Medicine, 79,* 555–560.

Littenberg, B. (1992). Technology assessment in medicine. *Academic Medicine, 67,* 424–428.

Osheroff, J., & Bankowitz, R. (1993). Physicians' use of computer software in answering clinical questions. *Bulletin of the Medical Library Association, 81,* 11–19.

Salton, G. (1992). The state of retrieval system evaluation. *Information Processing and Management, 28,* 441–449.

Salton, G., Buckley, C., & Smith, M. (1990). On the application of syntactic methodologies in automatic text analysis. *Information Processing and Management, 26,* 73–92.

Saracevic, T. (1975). Relevance: A review of and a framework for the thinking on the notion in information science. *Journal of the American Society for Information Science, 26,* 321–343.

Schamber, L., Eisenberg, M., & Nilan, M. (1990). A re-examination of relevance: Toward a dynamic, situational definition. *Information Processing and Management, 26,* 755–776.

Schwartz, S., & Griffin, T. (1986). Learning, feedback, and decision aids. In *Medical thinking: The psychology of medical judgment and decision making* (pp. 158–215). New York: Springer-Verlag.

Swanson, D. (1988). Historical note: Information retrieval and the future of an illusion. *Journal of the American Society for Information Science, 39,* 92–98.

Wiliamson, J., German, P., Weiss, R., Skinner, E., & Bowes, F. (1989). Health science information management and continuing education of physicians. *Annals of Internal Medicine, 110,* 151–160.