EVALUATION OF EVALUATION IN INFORMATION RETRIEVAL

Tefko Saracevic, PhD
School of Communication, Information and Library Studies,
Rutgers University
4 Huntington St.
New Brunswick, NJ 08903
saracevic@zodiac.rutgers.edu

ABSTRACT

Evaluation is a major force in research, development and applications related to information retrieval (IR). This paper is a critical and historical analysis of evaluations of IR systems and processes. Strengths and shortcomings of evaluation efforts and approaches are discussed, together with major challenges and questions. A limited comparison is made with evaluation in experts systems and Online Public Access Catalogs (OPACs). Evaluation is further analyzed in relation to the broad context and specific problems addressed. Levels of evaluation are identified and contrasted; most IR evaluations were concerned with the processing level, but others were conducted at the output, users and use, and social levels. A major problem is the isolation of evaluations at a given level. Issues related to systems under evaluation, and evaluation criteria, measures, measuring instruments, and methodologies are examined. A general point is also considered: IR is increasingly imbedded into many other applications, such as the Internet or digital libraries. Little evaluation in the traditional IR sense is undertaken in relation to these applications. The challenges are to integrate IR evaluations from different levels and to incorporate evaluation in new applications.

Introduction

Evaluation means assessing performance or value of a system, process (technique, procedure . . .), product, or policy. As such, evaluation is accepted as a critical necessity in science, technology, and many other areas, including social applications. What should form the basis of evaluation is often a difficult and vexing problem. Thus, in any area of evaluation much research is done on evaluation criteria, measures, methods, and related aspects. At times it is necessary to lean back and ask even more general questions, namely valuative questions about evaluation itself. This is such a paper.

SIGIR'95 Seattle CA USA© 1995 ACM 0-89791-714-6/95/07.$3.50

From the inception of IR systems a half century ago, evaluation was a major force in the progress of IR research and development (R&D). In this paper I wish to analyze the basic and historical aspects related to evaluation of IR systems and processes, assess the accomplishments and shortcomings of IR evaluation, raise critical issues and questions, and compare in a limited way IR evaluation with evaluation of related information systems, most notably expert systems and OPACs.

Context: the broad problem

In a most influential article Vannevar Bush, the head of U.S. scientific efforts during the Second World War, addressed the problem related to "the massive task of making more accessible the bewildering array of knowledge" in science and technology, and suggested the application of the emerging information technology as a solution to the task (Bush, 1945). Bush was concerned with the problem of 'information explosion,' the accelerated and exponential growth of records of knowledge in science and technology. Many others shared the concern and were enthusiastic about the suggested technological approach for solution. As a result, by the 1950s IR emerged, fueled by significant government funding, as the major approach in addressing information explosion, first in science and technology and later in all the areas of human endeavor.

Information explosion remained to this day the context of IR as the major problem addressed. IR is still being funded because of it. Over time, IR changed significantly, even the nature of our understanding of the information explosion phenomenon changed, but the basic problem orientation for IR remained constant. Thus, the ultimate evaluation question for IR would be: *How successful was and is information retrieval in resolving the problem of information explosion in the areas applied?* Clearly, this begs many related questions, among them: *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?*

These are very hard questions to address, because the issues are not just a matter of systems performance. After all, information explosion is a complex social, cognitive, cultural, and communication problem, and not just a technical problem. Thus, at its base IR is an activity oriented toward problems with the same (social, cognitive, communication) characteristics. Ultimately, IR exists as a social activity to link

and enable interaction between producers/authors of information (or more specifically, the texts, data, images, sounds . . . they produced) and users/readers of information. However, an indirect valuative answer may be derived.

Over decades, the results of IR R&D were widely applied. The IR ideas and experiments from the 1950's and 1960's resulted in development of massive IR databases, systems, services and networks in the 1970's and 1980's. A strong, successful and rich information industry evolved based on IR. Research from the 1980's and 1990's is filtering now into this industry. Much of the practice and work of information professionals was changed, revolving around IR. The information industry and IR are showing signs of maturity.

Thus, from a pragmatic and commercial point of view the systems and processes developed in IR became highly successful They are ubiquitous. IR is showing up everywhere, and not only in information industry. Among others, IR is applied to search and navigate through the Internet; IR is also the centerpiece of R&D on digital libraries.

But these positive and highly successful aspects of IR have a negative side as well. Being ubiquitous has a price. Some of the present IR applications or proposals came from people and places that are, to put it frankly, naive in IR, with little understanding of the complexity of the problems or the achievements in IR work. Consequently, many proposals for future IR R&D and many applications are steps backward, including some that are related to the Internet and digital libraries. They are immature and substandard. Unfortunately, there is little evaluation of many of these broader IR efforts.

Let me now draw some comparisons. The work on expert systems, coming from artificial intelligence (AI), emerged in mid 1960's It soon spread and gained in popularity. The problem addressed in expert system research was also related to knowledge. However, there is a difference: IR dealt with physical records of knowledge, while expert systems addressed a different kind of record: the cumulated knowledge in the minds of experts. Evaluation was not a strong consideration in expert system R&D, as it was in IR R&D. Great many expert systems and expert system shells were developed, but not that many survived In comparison to IR applications, the applications in expert systems were not as widespread or successful. They proved less robust and appropriate in their techniques then IR applications Crevier (1993) undertook a historical analysis of AI, including an evaluation of expert systems successes and failures; it would be interesting to undertake a parallel comparative evaluation that included IR.

Evaluation of libraries addressed a variety of aspects and problems, but, as mentioned, I will deal here only with OPACs because of their close relation to IR The work in application of information technology to libraries, labeled 'library automation', emerged in 1960's. Originally it addressed the problem of automation of internal library processes, most notably acquisition, cataloging, and circulation. These efforts had no connection with IR. However, beginning in the late 1970's library automation evolved to include the problems of access to and searching of library catalogs by users. OPACs were the result. OPACs addressed, as did IR systems, problems of records of knowledge, but a specific type of standardized records as found in library catalogs. Evaluation was not a strong priority in OPAC research and application; their evolution was not guided by widespread evaluation, as was IR's. Nevertheless, OPACs were successful in applications. They spread widely. OPACs have a strong IR component, but the IR community has only sporadically incorporated OPACs in their R&D. As with expert systems, it would be beneficial to undertake comparative evaluations of IR systems and OPACs

The specific problems: Nature of IR processes

When considering evaluation, we must starts with considering what problems and processes are being evaluated. Calvin Mooers coined the term 'information retrieval', and defined the specific problems to be addressed (Mooers, 1951):

> "Information retrieval is the name for the process or method whereby a prospective user of information is able to convert his need for information into an actual list of citations to documents in storage containing information useful to him . Information retrieval embraces the intellectual aspects of the description of information and its specification for search, and also whatever systems, techniques or machines that are employed to carry out the operation."

After half a century of evolution IR systems and processes have become highly sophisticated. Of the many changes and improvements probably the most significant is that IR systems now provide for a high degree of interaction, with all the accompanying implications and problems of human-computer interaction. This, of course, led to consideration of IR in broader information behavior (seeking and using) context, and not only in a narrow technical context. In addition, while much of IR still deals with citations, many IR systems also went way beyond citations, to cover texts , data, and images. However, the problems identified by Mooers are still at the base of IR. *How to provide a prospective user with useful information?* Or in contemporary terms: *How to provide users with effective access to and interaction with information, and enable them to effectively use information?* And for that objective:

1. How to organize information *intellectually?*
2. How to specify the search and interaction *intellectually?*
3. What systems and techniques to use for those processes?

Thus, the implication for IR evaluation is that the assessment of performance and value is still related to these questions.

The role of evaluation in IR

Concerns about evaluation of IR systems came shortly after the appearance of first design proposals and first prototypes. Kent et al. (1955) were the first to propose the criterion of relevance and the measures of precision and relevance (later renamed recall) for evaluation of IR systems; these became the staple of most IR evaluations to this day. Shortly after that, various U.S. government agencies began sponsoring IR evaluation efforts. The most famous of the pioneering IR evaluations were the Cranfield studies that begun in the late 1950s and ran till mid 1960's. Cranfield is the grand-daddy of IR evaluations, setting the role for evaluation in IR, and the tone and approach that is used in most evaluation studies to this day.

Starting with the tradition established by Cranfield and then by numerous evaluations under the umbrella of SMART (Salton, 1971), designs of IR systems, including proposals for new methods and algorithms for IR processes, on the one hand, and their evaluation on the other hand, became inseparable. Evaluation became central to R&D in IR to such an extent that new designs and proposals and their evaluation became one. Cranfield, SMART and other studies up to 1980 are discussed in the excellent book on IR evaluation, edited by Sparck-Jones (1981). An update is needed.

Similar centrality of evaluation was not the case in R&D related to expert systems or OPACs. Evaluation, as originally carried out in expert systems and OPACs, was restricted to a technical or engineering level, trying to determine whether the software in question worked at all, whether it was bug free (i.e. evaluation as to errors, faults, and failures), and what is the software reliability, robustness and the like. Of course, these are very important and difficult assessments, but they are also very restrictive.

Broader evaluation, i.e., broader then software concerns and levels, did not become and it still is not at the center of evaluation of expert systems, as shown in a comprehensive review and proposal for evaluation of expert systems by Kirani, Zualkerman, and Tsai (1994). In other words, expert systems were rarely, if ever, formally evaluated beyond the engineering level. Evidently, expert systems have serious problems stemming from the very base of their assumptions and design ideas (Crevier, 1993). I would hypothesize that this is because their evaluation was restricted to the engineering (software) level; little or no serious attention was paid to the broader evaluation, involving other levels (described below) related to users and use, fitness-of-use, and impact. The software clearly became better and better, but the systems were on the wrong track. There is a lesson here to learn for IR evaluation.

Approaches to evaluation

Many approaches could be used in evaluation. The choice of the approach depends on the intent of evaluation and in many ways it defines the type of results obtained. In addition, every approach has its limitations. It is hard to mix approaches. With this in mind I will analyze the major approaches used in evaluation to IR.

For obvious reasons the overwhelming majority of IR evaluations have used the system approach. This approach is used to such an extent that people using it often neglect to recognize or use important findings from other approaches. A while ago Churchman (1968) has pinpointed not only the strengths, but also the limitations and blind spots of the systems approach. *Are these recognized in IR evaluations?* The strengths are, but the blind spots are not.

A system can be considered a set of elements in interaction. A human-made system, such as an IR system, has an added aspect: it has certain objective(s). The elements, or the components, interact to perform certain functions, or processes, to achieve given objectives. But as everything else in systems, objectives appear in hierarchies. Furthermore, any system (IR systems included) exists in an environment or environments (which can also be thought of as systems), and interacts with its environments. It is difficult and even arbitrary to set the boundaries of a system. In evaluation of IR, as in evaluation of any system or process, these difficult questions arise that clearly affect the results: *Where does an IR system under evaluation begin? Where does it end? What are the boundaries? What to include? What to exclude?*

To evaluate a system means to ask questions about its performance in relation to given objective(s): *How well does the system (or a component thereof) perform that for which it was designed?* It follows that evaluation cannot be undertaken without explicitly or implicitly having some objective(s) in mind. That sounds simple and logical. *But, which objective of an IR system, in the hierarchy of objectives, should be addressed?*

The first dilemma and difficulty in evaluation are the selection of the level of objectives to address. Let me divide objectives, and thus evaluations, of a technical, computer-based system, such as an IR system or expert system, into six general classes or levels (of course, they are not mutually exclusive):

1. On the *engineering level* question of hardware and software performance are addressed, such as reliability, errors, failures faults, speed, integrity, maintainability, flexibility, etc. In IR, computational effectiveness and efficiency of given retrieval methods and algorithms are investigated

2. On the *input level* questions about the inputs to and contents of the system are investigated. In IR, questions about coverage in the designated area are asked.

3. On the *processing level* questions about the way the inputs are processed arise. In IR these include assessment of performance of algorithms, techniques, approaches, and the like.

4. On the *output level* questions about interactions with the system and obtained output(s) are addressed. In IR these include assessment of searching, interactions, feedback, given outputs, and so on.

5. On the *use and user level* questions of application to given problems and tasks are raised. These questions are asked in IR as well, but they also include questions about market and fitness-of-use.

6. On the *social level* issues of impact on the environment(s) occur. In IR effects on research, productivity, decision-making and the like, in a given area are asked.

Moreover, economic or efficiency questions can be asked and contrasted at each level. Evaluation on one level rarely, if ever answer questions from another. For instance, evaluations of engineering or processing aspects of IR systems say little about questions arising in evaluation of use. In real-life operations and applications of IR these levels are closely connected. In evaluations of IR systems they are not. As yet, we did not achieve comprehensive evaluation of an IR system on more than one level. Nor was this achieved in expert systems or OPACs. *This isolation of levels of evaluation could be considered a basic shortcoming of all IR evaluations.*

IR evaluations at different levels

In this section I would like to describe more specifically IR evaluations associated with different levels and approaches, and draw some general conclusions. Cited works are but examples of much larger literature.

Most of the efforts and literature in IR evaluations are on the processing level. Examples are large evaluation projects like Cranfield (Cleverdon, Mills & Keen, 1966), SMART (Salton, 1971, 1989), and TREC (Harman, 1995). Yet many evaluations were carried out also on the output level and the user and use level. Examples are evaluation studies by Fenichel (1981), Borgman (1989), Saracevic, Kantor, Chamis & Trivison, (1988), Haynes et al. (1990), Fidel (1991), Spink (1995) and many others. While results of these studies are interesting in themselves, they also have implication for design, but were never considered by the evaluations on the processing level. It says in effect, that studies concerned with processing and algorithms are not interested in studies on output and users and use levels, and vice versa. There were also a few studies on the social level, evaluating the impact of IR systems in an area, e.g., studies of impact of MEDLINE on clinical decision making (Lindberg et al., 1993), but these were also ignored by evaluations at other levels.

There are two further classes of relatively isolated evaluations. The first one deals with end-users. As end-user searching of IR systems expanded, there emerged a large body of studies evaluating end-user performance and use of IR systems (e.g., Meyer & Ruiz, 1990; others are summarized in Dalrymple & Roderer, 1994). This growing number of evaluations are taking quite different approaches, and asking quite different questions then studies so far reviewed. However, they represent real evaluations of real users and uses of IR, with direct implications for design.

The second large class of IR evaluations deals with markets, products, and services from information industry. They also exist in relative isolation of the other evaluations mentioned. These are evaluations done also on the output and user and use levels. However, they are using actual products and services on the market as the base and aim of evaluation. For instance, there were numerous studies evaluating different media, such as CD-ROMs, available for IR (e.g., Rapp et al. 1990). Information professionals have been constantly involved in such evaluations, and so were many database producers and services. (Such evaluations appear regularly in professional journals such as *Online, Online Review, Searcher, RQ, Database, Library Journal,* and others). Most of these concentrate on specific features of products and services, and in that they are limited, but they carry an enormous weight in the professional community. The research community should look at times what was evaluated and what was the conclusion in these market studies, for they also contain clear implications for design.

The point is, *there is much, much more to evaluation of IR than evaluation of a variety of algorithms and procedures.* In particular, it should be realized that evaluations of IR on the processing level though numerous, popular, and undeniably successful, are also restricted to its own level, and thus have a serious blind spot.

In a widely cited review, Dervin & Nilan (1986) contrasted the system-centered evaluation (roughly the first three levels identified above) with the user-centered evaluation (roughly the last three levels), and issued an impassioned call for a paradigm change or shift in IR evaluations from system- to user-centered evaluations. They put it in an either-or proposition. They were wrong, dead wrong. A paradigm change from one to another orientation in evaluation is not needed. *Both system- and user-centered evaluations are needed.* But, they should (and even must) work together and feed on each other to achieve a more comprehensive picture of IR performance and avoid dangerous blind spots (e.g., ala expert systems). If there is a paradigm shift, it should be toward cooperative efforts and mutual use of results between system- and user-centered evaluations The shift should be toward braking not making barriers.

Requirements for evaluation

When considering IR evaluation, it is important to consider the requirements needed for any evaluation. Evaluation of any system, IR systems and processes included, requires ALL of the following: (i) *a system*, or its representation: a prototype, product . ., together with a *process* (algorithm, simulation, and the like), e.g., in IR the test collection in TREC and associated processing under given algorithms/procedures; (ii) *a criterion or criteria* representing the objective(s) of the system, e.g., in IR relevance as criterion; (iii) *measure(s)* based on the criterion or criteria e.g. recall and precision; (iv) *a measuring instrument(s)* to register the measure e.g. relevance judgements by analysts in TREC; and (v) *methodology* for obtaining measurements and conducting evaluation e.g. the setup and procedures for TREC.

Research can be done on each of these requirements, as it has been done in IR. An excellent and specific list of requirements for IR evaluation was provided by Tague-Sutcliffe (1992), showing the complexity of any evaluation activity.

I am using these five requirements as a classification for further and more detailed discussion of evaluation in IR. I am also suggesting that these requirements can serve as a general framework for any discussion of evaluation activities in IR. Spark Jones (1995) has used a similar framework for an excellent analysis of TREC structure, conduct, and findings.

Systems and processes in IR evaluation

Majority of evaluations on the processing level (i.e., evaluations of algorithms and procedures) were done on databases assembled especially for that purpose. The Cranfield and various SMART collections were small. They were laboratory collections successfully affording control; they dominated IR evaluations for close to three decades. Much was learned from using them and in itself the state of IR evaluation made significant methodological progress (Salton, 1992). But in relation to IR as practiced they were toy collections, vastly removed from any reality. This was a serious deficiency. While many results were interesting and illuminating, they still have to be treated merely as hypotheses for industrial strength evaluations and applications.

TREC is an attempt to remedy that deficiency. Databases in TREC are large (over one million documents), approaching the size of many databases used in practice (e.g., MEDLINE has over 7 mill. records). However, TREC also has a highly unusual composition as to types and subjects of documents. *How does this effect the approaches tested and the very results obtained?* Nobody knows. This and other methodological problems related to TREC (discussed later) should be considered and their effects, if any, investigated. Nevertheless, TREC is a large step forward. Among the contributing factors are: its large size, large scale of processes/algorithms evaluated, diversity of approaches, stringency of requirements, and concurrence of many teams working at the same time affording comparison and even cooperation. In a short time TREC has exercised an enormous influence on the field, with more to come. TREC is determining IR evaluation for the 1990's and probably beyond. But on the flip side one has to ask the perennial nagging question: *Could all eggs in one basket be a danger as well? What are the pitfalls of the concentration on TREC as practically the sole vehicle for IR evaluation?*

SMART and TREC are used to evaluate algorithms and approaches (e.g., statistical, linguistic) used for organizing, searching, assembling, and/or retrieving outputs from the given databases. An impressive body of experience has been accumulated through these experiments As a result, approaches are becoming more sophisticated and mature. However, at least one vexing question should be raised. Every algorithm and every approach anywhere, IR included, is based on certain assumptions. As Cooper (1994) pointed out, the underlying assumptions and "trappings" of IR algorithms have a potential for serious effects. *How well are the assumptions for different methods understood and explained?* And conversely: *How much are they ignored? What are the traps?* These basic questions need evaluation.

Let me give an example about the danger of ignoring assumptions from the experiences with expert systems. The basic assumption underlying these systems is that human knowledge and expertise can be reduced to and expressed in rule-based (if . . . then) reasoning. The assumption was never investigated. One of the explanation of a restricted success of these systems, and an outright rejection in a number of domains, is that some human knowledge and expertise can be reduced to rule-based, but other is much more complex for such reductionism; in those cases expert systems don't work (Crevier, 1993). The assumption was not fully warranted

But the most serious criticism and limitation of the SMART and TREC evaluations lie elsewhere. These projects treat IR in a batch mode (as did Cranfield) and not in an interactive mode. Despite inclusion of relevance feedback techniques in evaluations and planned "interactions" in future TRECs, these are still static systems. They do not evaluate interactive IR. In contrast, most of IR in practice is interactive. IR cannot be even envisioned any more without interactivity - it is not a but THE major feature of IR. IR interactions are rich and complex; they have large effects on the process and results. Interactions in IR were extensively studied and modeled (Ingwersen, 1992). Yet interactivity plays practically no role in present large evaluation projects, TREC included. Interaction was addressed in some evaluations of OPACs (Robertson & Hancock-Beaulie, 1992), but has yet to be addressed in IR evaluations on the processing level. Admittedly, this issue was much debated. Furthermore, it is most difficult to incorporate interactions in evaluation. But still, *interaction must be addressed beyond debates in order for IR evaluation to catch up with reality.*

Evaluation studies oriented toward output and user and use levels employed as a rule existing, operational databases and settings to make observations (as opposed to evaluations on the processing level done under laboratory conditions). The approach, of course, added realism of observations, but reduced possibilities for control. Some control was exercised by controlling the population and conditions of observations in these studies. In addition several of these evaluations did include interactivity as a major aspect investigated. Unfortunately, since there was such a diversity of systems and processes involved, scientific cumulation of findings is difficult and rare.

Criteria in IR evaluation

As mentioned, Kent et al. (1955) proposed relevance as the basic (and sole) criterion for IR evaluation. It stuck. Relevance remained the only evaluation criterion for overwhelming majority of studies on the processing level Other criteria were proposed, among them utility and search length, but they did not stick. Thus, the studies such as Cranfield, SMART, TREC, and many others, revolved around the phenomenon of relevance. Selection of this single criterion had an enormous and rarely thought about impact on IR evaluations. On the positive side it made evaluation studies focused, comparable, and easier. It also lifted them out of a dangerous isolation of studies on the engineering level, because use was implied. However, selection of relevance also had its drawbacks. Relevance is a complex human cognitive and social phenomenon; as many such phenomena it is elusive, and messy as well. Because of the selection as the criterion for IR evaluation, relevance also became a subject of intensive study of its own in information science (Saracevic, 1975, Schamber, 1994), to the point that it is considered among the basic phenomena of the field. Findings include evidence of a variety of variables that effect its behavior, e.g., there are considerable individual differences in relevance assessment by people; relevance is assessed by people in gradations, i.e., it is not a binary yes-no decision; it also heavily depends on circumstances, context, and so on. Iinterestingly and unfortunately, the findings from the studies of relevance on the one hand, and the use of relevance as a criterion in IR evaluations, on the other hand, have no connection. A lot is assumed in IR evaluations by use of relevance as the sole criterion. *How justifiable are these assumptions?*

The mentioned studies on output and user and use levels used a number of criteria besides relevance These include criteria related to utility, success, completeness, satisfaction, worth, value, time, cost, and so on. By using these additional criteria a much more complete picture is gained of user reactions and behaviors. In addition, such criteria seem to be more suitable for evaluation of IR interactions.

The market evaluations used some of these same expanded criteria for evaluation, but went heavily in favor of fitness-of-use criteria of quality, as common in Total Quality Movement (TQM), and pragmatic and cost-effectiveness criteria common in business and industry.

A major challenge for IR evaluation efforts is to apply these various criteria in some unified manner, to obtain a more comprehensive picture than any one provides alone Of course, it is so much easier to use one criterion for evaluation then many. It is also more restrictive.

Criteria for expert system evaluation were similar to those listed above under the engineering level of evaluation. Criteria for OPEC evaluation were similar to the multiple criteria used in IR studies on output and user and use levels. Interestingly, in neither of these areas can be found concerns, studies and intensive debates about the criteria used, as in IR where relevance was a major area of study and debate to this day

Measures in IR evaluation

Following selection of relevance as criterion, precision and recall became the preferred pair of measures of IR evaluation studies on the processing level. (Precision is the ratio of relevant items retrieved to all retrieved items, or the probability given that an item is retrieved it will be relevant. Recall is the ratio of relevant items retrieved to all relevant items in a file, or the probability given that an item is relevant it will be retrieved). In Cranfield, SMART and TREC they are the sole measures. There were many proposals for a unified measure (as opposed to a pair of measures) from Swets (1963) to Wilbur (1992). However, these remained just proposals Their applicability has yet to be evaluated.

Given a relevance assessed output, precision can be directly and easily derived. Recall can not. This is because recall depends not only on what was retrieved, but also on what was not retrieved, i.e., on what was missed. In this sense recall is a metaphysical measure: how does one know what is missed when one does not know that it is missed? Thus, establishing a set to calculate recall presents methodological challenges Many methodological "tricks" were devised to be able to estimate recall. In toy databases all documents were assessed as to relevance for given requests beforehand. In large databases, as TREC, this is not possible. The method of polling (combining) of all the outputs for the same request is used to establish relative recall. This may be fine for comparisons.

Butindependent of such methodological issues, a large, hidden and troubling assumption underlies any and all use of recall as a measure. The assumption is that there exists one and only one relevant set in a database as an answer to a request. This is a tall assumption, operationally not warranted, particularly in large databases. In several studies (e.g. Saracevic et al 1988; Haynes at al. 1990)a relatively small overlap was found among relevant items retrieved by different users or searchers

searching for the same request. People select different terms and strategies for searching, and consequently retrieve different sets of items, both relevant and not relevant. Yet they find the outputs satisfactory. In fact, in a large database, such as MEDLINE, many relevant sets may exist and satisfy user needs for a given request. *What warrants the assumption that there is only ONE relevant set out there for any given request?*

In TREC different algorithms and approaches tend to retrieve differing subsets of documents for the same request with relatively little overlap, yet their recall figures are not always that different. The whole issue of the nature and extent of overlap in TREC (and elsewhere) should be investigated. Clearly, if one abandons the assumption of only a single relevant set per request, the whole use of the measure of recall must re-evaluated. Recall may go out of the window.

Evaluations on other (output, users and use, and social) levels often tend to use semantic differentials or Likert scales as measures. As ordinal measures they have well-known properties and limitations. Since many criteria and measures were used, the perennial questions remain: *Which measure(s) to use? How do the measures compare? Which one is 'better' in given circumstances?* A study by Su (1992) is an example of the regrettably small number of studies that have addressed these questions. Many more are needed.

Measuring instruments in IR evaluation

Measures need some means, i.e., instruments, to be registered. For the reviewed measures people were used as measuring instruments. They judged relevance of retrieved items, or for ther measures, they entered their assessment on a measuring scale directly. Of course, this brings immediately several well-known questions: *Who should b, the judges? How reliable are their assessments? What effects their judgment? How do they effect the results?* These questions are recognized as serious problems in all evaluations where people are used as instruments, not only in IR. Considerable research is done in these areas. A generation ago there were a few large studies in IR that addressed such questions, but as we can see from the review by Schamber (1994), very few if any studies on this topic are conducted now. In IR these questions and associated problems are sometimes acknowledged, but mostly lamented upon and then swept quietly under the rug.

The approach taken by large IR evaluations (Cranfield, SMART, TREC) is to take assessments by user surrogates (i.e., by people assumed to represent users in some way) as is without any further questions. This affords uncomplicated measurements, and easy comparison, with bias, if any, being equal for every process evaluated. The assumption is this: if relevance is the criterion and precision and recall the corresponding measures, then somebody has to judge or establish relevance; we will assume that that somebody made a reasonable judgement; then any possible effects do not matter - they are the same for everything evaluated. It is an uneasy assumption, often questioned. Spark-Jones (1995) has asked some pointed questions about analysts used in TREC for judgement of relevance. *How do the relevance judges effect the result?* Who knows?

Methods in IR evaluation

Methods refer to the design, manner, means and procedures used to get and analyze evaluation results. By themselves methods are evaluated for their validity, reliability, appropriateness and related criteria. In IR there is a long tradition in examining, and challenging the methodologies used in evaluation. For instance, Cranfield results were seriously challenged by Swanson (1971) because of methodological faults. Cooper (1968) was among the first in the long line of authors who raised questions of integrity of the whole IR evaluation enterprise where relevance judgments were used. After a large evaluation of full text retrieval in legal research by Blair & Maron (1985), Salton (1986) raised a number of methodological issues bringing the whole project into question, and Blair & Maron (1990) retorted explaining and justifying their methodology for evaluation. Many troubling methodological issues have been brought out by articles in the book by Sparck-Jones (1980); similar issues were brought out again and more recently by articles in Harman (1992), and by Sparck-Jones (1995). And so it goes. Methodology was a troubling and challenging part of IR evaluations. It still is.

Methodological issues, critiques and questions, particularly as related to validity and reliability, fall in these general categories: (i) Collection (data). *How are the items (e.g., documents) selected that are used in evaluation? How appropriate (homogeneous, representative . . .) are they? What size is the collection? How are the items in the collection treated?* (ii) Requests: *How are they generated? How representative, appropriate . . . are they? How are they treated?* (iii) Searching: *How is it conducted? What procedures are used? How realistic?* (iv) Results: *How are they obtained? Who are the judges? How is the judging conducted? How realistic? How is the feedback handled, if any?* (v) Analysis: *How is it performed? What statistical or other tools are used? What comparisons are made and how?* (vi) Interpretation, generalization: *What are the conclusions? Are they warranted on basis of results? How generalizable are the findings?*

As can be seen in the literature, each successive generation of IR evaluations is raising these questions anew. As long as there will be evaluations such questions have to be raised and dealt with.Conversely, *how dangerous is it not to consider them?*

Conclusions

Evaluation is an integral and vital part of IR. This paper is an attempt to evaluate a broad range of evaluations in IR, identify

major strengths and shortcomings of evaluations, raise questions that ought to be considered in evaluation, and provide some comparisons with evaluation of related information systems. IR itself and IR evaluation has a long and proud history. In many respects they were successful. However, many issues and problems in evaluation are reappearing in historical cycles. Great many significant valuative problems need to be addressed.

I consider that among the major problems is isolation of evaluations of IR at different levels. IR is evaluated as to the algorithms on one level, as to the users and uses on another, as to the market products/service on still another, and as to the social impacts on yet another. Each of these serves a given purpose. However, the results are relevant and applicable at more then one level. For instance, design decisions can be guided, and justified by results from any level. The issue is not which level of evaluation is "better" or "best." *The issue and challenge for any and all IR evaluations are broadening of approaches and getting out of the isolation and blind spots of single level, narrow evaluations.*

A good example is the obvious need to integrate evaluations on the processing level (i.e., evaluations of IR algorithms and procedures) with evaluations on the output and user and use levels, involving interaction and user behaviors. This is to keep evaluations realistic from both ends. A related question is: *How can interaction be ignored in IR evaluation at any level?*

Let me end the conclusions with some larger issues. Increasingly, IR applications can be found outside the area of IR proper, i.e , outside the traditional IR R&D, the resulting applications, and the related evaluations. The Internet is one example, digital libraries another. In those applications IR techniques are being applied or adapted similarly as those from AI. The metaphor is "like raisins in raisin bread." They are imbedded in larger applications, but they are still central and vital, for "there is no raisin bread without raisins." In those applications, evaluation as IR knows it is absent. While evaluation is talked about (e.g., every digital library project has some evaluation included) it is not central, it does not guide and even less govern the activity as it did in traditional IR. Evaluation, if carried out at all, is limited mostly to the engineering or software level, the same as it was in expert systems. Consequently, the same dangers loom. The software, systems, and networks may be evaluated as working well on their own level, but the real problems are not at that level and the effects may be negative elsewhere. As to the output, users and use many of these applications are shown to be well received, but they have also been shown among others to be frustrating, unproductive, trivial, wasteful, expensive, unreliable, unpredictable, and hard to use. On the social level they do have an impact - increasingly so. The impact is yet to be figured out, including the impact on traditional institutions such as business, libraries, education . . , and the impact on individuals. The social calculus is different from the engineering and systems one

The most significant and the most difficult valuative question for IR and for all these IR imbedded applications will be of this order: *How does all this information, and associated information technology and information systems effect our work, leisure, society, culture? How do IR and related applications reorder life?* Presently, we can hardly envision either the answers, or even the methods for trying to get at them. Nevertheless, that should not prevent us from thinking about these questions, and having such fundamental issues constantly in the back of our minds when engaged with nuts and bolts of everyday research and other work related to IR

REFERENCES

Blair, D.C & Maron, M.E. (1985). An evaluation of retrieval effectiveness for a full-text document-retrieval system. *Communications of the ACM, 28,* (3), 289-299.

Blair, D.C. & Maron, M.E (1990). Full text information retrieval: Further analysis and clarification. *Information Processing & Management, 26,* (3), 437-447.

Borgman, C.L. (1989). All users of information retrieval systems are not created equal: An exploration into individual differences. *Information Processing & Management, 25,* (3), 237-252.

Bush, V (1945). As we may think. *Atlantic Monthly, 176,* (1), 101-108.

Churchman, C.W. (1968). *The systems approach.* New York. Delta.

Cleverdon, C.W., Mills, J. & Keen, E.M. (1966). *An inquiry in testing of information retrieval systems.* (2 vols.). Cranfield, U.K.: Aslib Cranfield Research Project, College of Aeronautics.

Cooper, W.S. (1968). Expected search length: A single measure of retrieval effectiveness based on the weak ordering action of retrieval systems. *American Documentation, 19,* (1), 30-41.

Cooper, W.S. (1994). The formalism of probability theory in IR: A foundation or an encumbrance? *Proceedings of the 17th Annual International ACM-SIGIR Conference.* London: Springer-Verlag. pp 242-245.

Crevier, D. (1993). *AI: the tumultuous history of the search for artificial intelligence.* New York: Basic Books.

Dalrymple, P.W & Roderer, N. K. (1994). Database access systems. In Williams, M. E. (Ed.) *Annual Review of Information Science and Technology, vol. 29,* (pp. 137-178).

Medford, NJ: Learned Information.

Dervin, B. & Nilan, M.S. (1986). Information needs and use. In Williams, M. E. (Ed.) *Annual Review of Information Science and Technology, vol. 21,* (pp.3-33). White Plains, NY: Knowledge Industry.

Fenichel, C.H. (1981). Online searching: measures that discriminate among users with different types of experience. *Journal of the American Society for Information Science, 32,* (1), 23-32.

Fidel, R. (1991) Searchers' selection of search keys. (3 parts). *Journal of the American Society for Information Science, 42,* (7), 490-527.

Harman, D. (Ed.). (1995). The second Text Retrieval Conference - TREC 2. *Information Processing & Management, 31,* (3), Special issue.

Haynes, R.B.., McKibbon, K.A., Ryan, N., Fitzgerald, D., & Ramsden, M.F. (1990). Online access to MEDLINE in clinical setting: A study of use and usefulness. *Annals of Internal Medicine, 112,* 78-84.

Ingwersen, P. (1992). *Information retrieval interaction.* London: Taylor Graham.

Kent, A., Berry, M., Leuhrs, F.U. & Perry, J.W. (1955). Machine literature searching VIII. Operational criteria for designing information retrieval systems. *American Documentation, 6,* (2), 93-101.

Lindberg, D.A.B., Siegel, E.R., Rapp, B.A., Wallingford, K.T., Wilson, S.R. (1993). Use of MEDLINE by physicians for clinical problem solving. *JAMA, The Journal of the American Medical Association, 269,* (24), 3124-3129.

Meyer, D. E. & Ruiz, D. (1990). End-user selection of databases. 3 parts. *Database, 13,* part 1: (4), 21-29, part 2: (4), 35-42, part 3: (5), 65-67.

Mooers, C.N. (1951). Zatocoding applied to mechanical organization of knowledge. *American Documentation, 2,* 20-32.

Rapp, B.A., Siegel, E. R., Woodsmall, R., Lyon-Hartmann, B. (1990). Evaluating MEDLINE on CD-ROM: An overview of field tests in library and clinical settings. *Online Review, 14,* (3), 172-186

Robertson, S.E. & Hancock-Beaulieu, M.M. (1992). On the evaluation of the IR systems. *Information Processing & Management, 28,* (4), 457-466.

Salton, G. (1971). *The SMART retrieval systems: Experiments in automatic document processing.* Englewood Cliffs, NJ. Prentice-Hall.

Salton, G. (1986). Another look at automatic text-retrieval systems. *Communications of the ACM, 29,* (7), 648-656.

Salton, G. (1989). *Automatic text processing: The transformation, analysis and retrieval of information by computer.* Reading, MA: Addison-Wesley.

Salton, G. (1992). The state of retrieval system evaluation. *Information Processing & Management, 28,* (4), 441-449.

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,* (6), 321-343.

Saracevic, T., Kantor, P. Chamis, A.Y., & Trivison, D. (1988). A study of information seeking and retrieving. (3 parts). *Journal of the American Society for Information Science, 39,* (3), 161-216.

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

Sparck Jones, K. (Ed.) (1981). *Information retrieval experiment* London: Butterworths.

Spark Jones, K. (1995). Reflections on TREC. *Information Processing & Management, 31,*, (3),.

Spink, A. (1995). Term relevance feedback and mediated database searching: Implications for information retrieval practice and system design. *Information Processing & Management, 31,* (2), 161-171.

Su. L. T. (1992). Evaluation measures for interactive information retrieval. *Information Processing & Management, 28,* (4), 503-516.

Swanson, D.R. (1971). Some unexplained aspects of the Cranfield test of indexing performance factors. *Library Quarterly, 41,* (3), 223-228.

Swets, J.A. (1963). Information retrieval systems. *Science, 141,* (1), 245-250.

Tague-Sutcliffe (1992). The pragmatics of information retrieval experimentation, revisited. *Information Processing & Management, 28,* (2), 467-490.

Wilbur, W.J. (1992).An information measure of retrieval performance *Information Systems, 17,* (4), 283-298.