EVALUATING INFORMATION RETRIEVAL SYSTEMS UNDER THE CHALLENGES OF INTERACTION AND MULTIDIMENSIONAL DYNAMIC RELEVANCE

Jaana Kekäläinen & Kalervo Järvelin
University of Tampere
Department of Information Studies
FIN-33014 University of Tampere, FINLAND
Email: {jaana.kekalainen, kalervo.jarvelin}@uta.fi

Published in: Harry Bruce, Raya Fidel, Peter Ingwersen, and Pertti Vakkari (Eds.) *Proceedings of the 4th CoLIS Conference*. Greenwood Village, CO: Libraries Unlimited, pp. 253-270.

ABSTRACT

The Laboratory Model of information retrieval (IR) evaluation has been challenged by progress in research related to relevance and information seeking as well as by the growing need for accounting for interaction in evaluation. Real human users introduce non-binary, subjective and dynamic relevance judgments into IR processes and affect these processes. Therefore the traditional evaluation based on the Laboratory Model is challenged for its (lack of) realism. This paper examines the rationale of evaluating the IR algorithms, the status of the traditional evaluation, and the applicability of the proposed novel evaluation methods and measures. It further points out research problems requiring attention for further advances in the area. The Laboratory Model is found limited but still useful for the specific tasks it fulfills in the development of IR algorithms.

1. INTRODUCTION

The Laboratory Model of information retrieval (IR) evaluation has its origins in the Cranfield II project (Cleverdon, 1967). It is the paradigm of the Computer Science oriented IR research, seeking to develop ever better IR algorithms and systems. In recent years, the TREC conferences (Voorhees & Harman, 2001) have been the major forum for research based on the Laboratory Model. An essential component in evaluation based on the Model is a test collection consisting of a document database, a set of fairly well defined requests, and a set of (typically binary) relevance assessments identifying the documents that are topically relevant to

each request. IR algorithms are evaluated for their ability of finding the relevant documents. The test results are typically expressed in terms of recall and precision.

The Laboratory Model has been challenged by progress in research related to relevance and information seeking as well as by the growing need for accounting for interaction in evaluation. Work in analyzing the concept of relevance has resulted in identifying higher-order relevances, such as cognitive relevance and situational relevance, in addition to algorithmic and topical relevance (Borlund, 2000; Cosijn & Ingwersen, 2000; Saracevic, 1996; 1997; Schamber, Eisenberg & Nilan, 1990). Real human users of IR systems introduce non-binary, subjective and dynamic relevance judgments into IR processes, which affect the processes directly. In this sense relevance is multidimensional and cannot be derived from any single relevance criterion (Cuadra & Katter, 1967; Barry, 1994; Saracevic, 1975, 1997; Schamber, 1994). By *higher-order relevance* we refer to this subjective whole which besides topicality includes, *among others,* situational, cognitive and affective relevance.

Theoretical and empirical work in information seeking and retrieval (Belkin, 1993; Byström & Järvelin, 1995; Ellis & Haugan, 1997; Ingwersen, 1996; Kuhlthau, 1993; Schamber, 1994; Vakkari, 2001; Wilson, 1999) suggests that IR is but one means of information seeking which takes place in a context determined by, e.g., a person's task, its phase, and situation. IR tactics and relevance assessments are affected by the stages of task performance. Also some user-oriented research in IR, e.g., by Bates (1989; 1990) points out the variety of strategies users might use in accessing information, topical retrieval being only one.

Because of these empirical findings and theoretical arguments, the traditional Laboratory Model of IR evaluation is challenged for its (lack of) realism. Harter and Hert (1997) give a detailed analysis of the evaluation of IR systems and the criticism towards traditional IR experiments. There are proposals (Borlund, 2000; Hersh & Over, 2000) concerning how IR evaluation should be done realistically and at the same time retaining as much control as possible. There also is empirical work (e.g., Vakkari, 2001) tracing interactive information seeking and IR processes providing models and methods for IR evaluation. Developers of IR algorithms should therefore consider how the algorithms are to be evaluated in a valid way from now on.

This paper examines the status of the traditional methods and measures, and the rationale of evaluating IR algorithms in an interactive system environment. It proposes research problems

requiring attention for both accusers and defenders of the Laboratory Model. It further seeks to combine the traditional experimenting with users and interaction. The main focus of the discussion is textual IR.

2. EVALUATING THE ALGORITHMIC COMPONENTS

The Laboratory Model is depicted in Figure 1. In this view the IR system consists of a database, algorithms, requests, stored relevance assessments. The system components are represented in the middle and the evaluation components on top, left and bottom in the lightly shaded area. The main thrust of the research has been on document and request representation and the matching methods of these representations. Only recently, in the Interactive Track of TREC (see Over, 1997), have users been involved (the darkly shaded area). Even so, the systems still have been evaluated on the basis of how the users are able to find documents deemed relevant in the test collection.

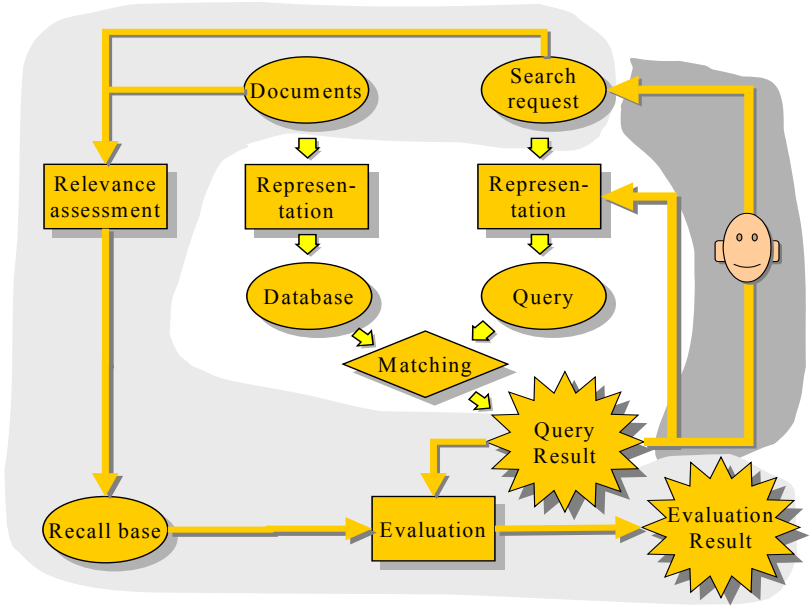

Figure 1. The Laboratory Model schematized.

In this view, real users and tasks are not seen as necessary. Test requests typically are quite well-defined requests with verbose descriptions that give the algorithms much more data to work with for query construction than typical real life IR situations do. Relevance is taken as topical, but factual features (based on structural data items, like author names and other bibliographic features) could be included. Relevance is static, between a request and a document as seen by an assessor. The assessments are independent of each other (i.e., no learning effects, no inferences across documents) and there are no saturation effects (i.e., in principle the

assessors do not get tired of, or mind, repetition). The assessors do not know in which order the documents would be retrieved so they cannot do otherwise or properly model user saturation.

The rationale of evaluating the algorithmic components consists of the goals, scope and justifications of the evaluation approach. The *goal* of research is to develop algorithms to identify and rank a number of topically relevant documents for presentation, given a topical request. Research is based on constructing novel algorithms and on comparing their performance with each other, seeking ways of improving them. On the theoretical side, the goals include the analysis of basic problems of IR (e.g., the vocabulary problem, document and query representation and matching) and the development of theories and methods for attacking them.

The *scope* of experiments is characterized in terms of types of experiments, test collections and requests as well as performance measures. The experiments mainly are batch-mode experiments. Therefore each algorithm is evaluated by running a test set of queries, measuring its performance for individual queries and averaging over the query set. Some recent efforts seek to focus on interactive retrieval with a human subject, the TREC interactive track being predominant. The major modern test collections are news document collections. The major performance measures are recall and precision.

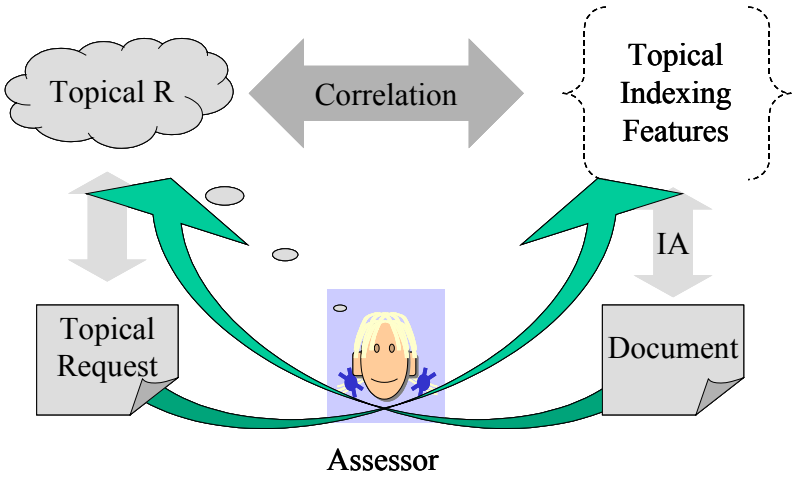

Figure 2. Justification of the Laboratory Model.

The *justifications* of the Model may be discussed in terms of Figure 2. The main strength is that words and other character strings from texts, when distilled as indexing features by an indexing algorithm, correlate, with fair probability, to the topical content of the documents

they represent and to the queries which they match (save for problems of homography). When a test user (or algorithm) processes a topical request, it is possible to predict, with fair probability, which indexing features should be considered (save for problems of synonymy, paraphrases). Because the topical request also suggests the topical relevance criteria, there is a fair correlation (clearly better than random) between the indexing features of matching documents and a positive relevance judgment. Indexing features correlate to meaning in the topical sense.

3. THE STATUS OF TRADITIONAL EVALUATION APPROACHES: OBJECTIONS AND RESPONSES

Objection 1. Lack of users and tasks.

There is no real user or task or situation involved, the Model is essentially based on "objective" assessors. Real IR is a subtask in task-based information seeking, and is thus affected by the latter (Kuhlthau, 1993; Vakkari, 2001; Wilson, 1999).

Response: True, but users and tasks are not needed for testing the algorithms for the limited goal they are intended for: retrieval and ranking of topical documents or their parts.

Objection 2. Lack of interaction and dynamic requests.

There is no real interaction and dynamics; it is essentially a batch mode evaluation. Real interaction involves user learning, problem redefinition and changing relevance criteria. (Beaulieu, Robertson & Rasmussen, 1996; Borlund, 2000; Vakkari, 2001). Test requests in interactive experiments are too rigid for the test users, and all documents that the users consider relevant are not deemed as relevant by the assessors, nor do the users accept all documents deemed relevant by the assessors.

Response: True, but not needed because all system activities in the interaction may be seen as individual retrieval tasks to be served well as such. Complex dynamic interaction is a sequence of simple topical interactions and thus good one-shot performance by an algorithm should be rewarded in evaluation. Changes in the user's understanding of his information need and relevance should affect the consequent request and query. Although a user is likely to modify his/her request and relevance criteria in subsequent interactions, it has not been shown that this should affect the design of the retrieval algorithm.

Objection 3. Lack of tactical variability.

The only tactic of interest is a batch mode topical (or question-answering) request while people in real life approach information in many different ways, including, e.g., bibliographic and other structured attributes or links (Bates, 1989).

Response: True, tactics deserve more attention, but the model is no hindrance for this.

Objection 4. Lack of ambiguity.

The requests are only well-defined requests, which do not correctly reflect all kinds of real-life requests; they are too well-specified and wordy for that (Ingwersen & Willett, 1995; Sparck Jones, 1995). They do not reflect the users' tasks or situations (Borlund, 2000).

Response: True, typical test requests are too well specified. This should be looked at and can be done within the model. For example, the different parts of TREC topics allow to investigate the effects of different query lengths (Voorhees & Harman 2001). Otherwise some IR models developed within the Laboratory Model explicitly tackle uncertainty (van Rijsbergen & Lalmas, 1996).

Objection 5. Lack of user-oriented relevance.

The tests are based on algorithmic and topical relevance, which are unable to take into account the user's situation, tasks, or state of knowledge. Relevance assessments also are stable – far from real-life. (Beaulieu & al., 1996; Borlund, 2000; Cosijn & Ingwersen, 2000.)

Response: True, but not needed for testing the algorithms for the limited task they are intended for. Higher-order relevance is out of scope of the framework unless explicated as request and document features to be processed – the algorithms do not read the users' minds. The heart of IR is matching explicit representations of documents and requests. The machine cannot do better if not designed to do so, which would require explication of higher-order relevance features both theoretically and in practice.

Objection 6. Lack of variety in collections.

The test collections, albeit nowadays large, are structurally simple (mainly unstructured text) and topically narrow (mainly news domain). The test documents mostly lack interesting internal structure that some real-life collections do have (e.g., field structure, XML, citations).

Response: True and should be looked at, the model is no hindrance. There is recent work in this direction (e.g., TREC Web Track, Hawking & al., 2000).

Objection 7. Document independence and overlap.

There are unrealistic assumptions regarding document independence (some may be relevant only if juxtaposed) and user saturation (repeated reproduction of very similar "relevant" information results in irrelevance) (Robertson, 1977).

Response: True, but the assumptions are a necessity since the relevance assessment stage is not informed about the possible combinations of documents retrieved by a query. No-one has been able to formalize the process of arguing across documents, such a task remains entirely in the user's domain.

Objection 8. Insufficiency of recall and precision.

Recall and precision are insufficient as evaluation measures, the former being system-oriented and often irrelevant to the user. They do not handle non-binary relevance (Borlund, 2000). They do not describe users' success in information problem explication.

Response: Recall and precision are major effectiveness measures for the limited retrieval goal. They reflect the kind of relevance that was used in the assessments be it topical or higher. Recall and precision may also be generalized to handle non-binary relevance assessments, as shown by Kekäläinen and Järvelin (2002). In the TREC interactive track, *instance precision and recall* break document level in relevance judgments. The user-system pairs are rewarded for retrieving distinct instances of answers rather than multiple overlapping documents (Hersh & Over, 2000).

Objection 9. Heavy averaging.

Many experiments are based on heavy averaging over sets of query results, perhaps never looking at the performance differences at the individual query level (Hull, 1996), or individual documents / requests.

Response: Often true, but not a limitation of the model. It may be a limitation of the IR evaluation culture.

Objection 10. Just document retrieval.

IR is just document retrieval with little, if any, attention to document/information presentation for use.

Response: True, but document retrieval is a genuine task in information access and deserves attention. Clearly, other stages of information access and use should be examined as well. There is recent relevant work in Question Answering, e.g., in TREC (Voorhees, 2001b), and Information Extraction, e.g., in the Message Understanding Conferences (Gaizauskas & Wilks, 1998).

Objection 11. IR is tested in isolation.

In the Laboratory Model IR is tested in isolation, not as part of a larger system or as one alternative of information seeking.

Response: True, but if IR systems are embedded in a larger context, the variables of interest may turn out to be uncontrollable, or effects may be unobservable.

A Concession

The Laboratory Model provides a controllable setting, and repeatable and generalizable results. Admittedly this is building and testing retrieval engines in isolation. There is no guarantee of solving any real life problems unless informed by information seeking studies. The approach may have a distorted view on information access problems, and may miss some important ones.

4. TWO BROADER EVALUATION SCENARIOS

We shall first look at recent conceptions of relevance and then assess novel user-oriented measures. We consider for each of these, which problems, if any, of the Laboratory Model they attack and whether they are successful in this. Finally, we look at two IR evaluation scenarios to which the proposals and their assessment lead.

4.1. Higher-order relevance

The concept of relevance has been a difficult problem in Information Science throughout the years. Schamber (1994; & al., 1990) have argued that relevance is a *multidimensional* and *dynamic* phenomenon. Several recent studies have focused on *factors* affecting relevance judgements, and *dimensions* (or criteria) of relevance (Barry, 1994; Barry & Schamber, 1998; Bateman, 1998; Borlund, 2000; Cosijn & Ingwersen, 2000; Vakkari & Hakala, 2000; Wang & Soergel, 1998). Others have argued that there are various *kinds* of relevance (algorithmic,

topical, cognitive, situational, motivational relevance; Saracevic, 1996). Elements of these typologies can be traced back to, e.g., Cuadra and Katter (1967) and Wilson (1973).

In many studies topicality is seen as the basic criterion for relevance – a necessary but not sufficient condition (Burgin, 1992; Cooper, 1971; Froehlich, 1994). The other criteria used in real users' relevance assessments are tied either to the state of knowledge or situation of the user. By higher-order relevance we refer to relevance which is not solely based on topicality but *also the other criteria*.

One of the main problems of the Laboratory Model was seen to be its lack of user-oriented relevance, thus sacrificing realism to control. Below we consider situational relevance as an example of higher-order relevance. Cosijn and Ingwersen (2000) define situational relevance as a relation between the perceived situation, work-task or problem at hand and the information objects under assessment. Incorporation of higher-order relevance into evaluation also brings interaction, dynamic requests/needs and tasks into evaluation thereby improving the realism of evaluation.

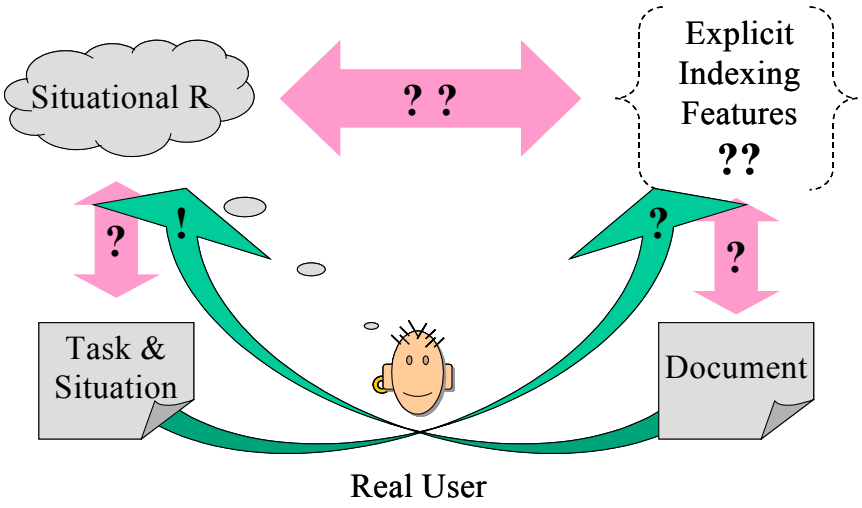

Figure 3. Situational relevance in retrieval

Some of the problems of using higher-order relevance in IR systems, or their design, may be discussed in terms of Figure 3. A real user, being thrown into a situation, may well be able to recognize a relevant document once presented (therefore the exclamation mark). However, he may have difficulty in discussing the relevance criteria of the task and situation. Further, he

certainly has difficulty in expressing a request and formulating a query to the IR system, at least anything other than topical *as long as text is concerned* (but save for bibliographic fields etc., if available, as discussed below), because current systems do not provide for anything else. The system designer probably never had the slightest idea of other than explicit topical indexing features, because there is no known pattern of situational indexing features that are explicit in text – the computer does not handle implicit features – and useful to users (see Cooper, 1971). Therefore the available indexing features may not correlate to the situational relevance criteria, which the user did not express, save for one thing: topical relevance heavily correlates to situational as shown by Burgin (1992), and Vakkari and Hakala (2000).

As soon as one moves away from collections consisting of plain text toward structured documents, the possibilities of applying situational relevance become better. For example, bibliographic data of a document (e.g., author, journal, year), its references and citations received may be useful a clues for situational relevance – and people actually use such criteria in relevance assessments (Barry, 1994; Schamber, 1994; Barry and Schamber, 1998). It is possible to compile lists of author names, institution names, journal and conference names in ranked order to signal relevance from a work-task or personal point-of-view. Also documents citing, or cited by, known situationally relevant documents might be ranked higher than documents lacking these explicit features. However, the person-to-person and task-to-task variation in such criteria may cause management problems – even if the criteria are explicit, they may be very private.

We therefore conclude that although it is easy to admit the realism of the higher-order relevance criteria, serving them by system features is a challenge because of lack of understanding on the variability of the criteria and on the combination of evidence on the criteria. There neither is much sense in evaluating the current algorithms' performance on criteria, which they completely ignore (save for correlation of topical and higher-order relevance).

4.2. Novel performance measures

Relevance is a *multilevel* phenomenon, i.e., some documents are more relevant than others to an information need of a user. Multiple *degrees* of relevance and their expression have been studied in laboratory settings (Cuadra & Katter, 1967; Rees & Schultz, 1967; Tang, Shaw & Vevea, 1999) as well as in field studies of information seeking and retrieval (Vakkari & Hakala, 2000). The former groups experimented with multiple point rating scales (from two to

eleven). Tang, Shaw and Vevea found that a seven-point scale for relevance assessments is optimal in terms of the assessors' confidence in their assessments.

However, until 2000 in the practise of IR evaluation, the binary scale has been the norm. This is unfortunate since it does not allow testing whether some IR methods are better than others at a particular degree of relevance. (Borlund, 2000) and Schamber (1994; referring to Cuadra & Katter, 1967 and Rees & Schultz, 1967) note that measures like recall and precision based on binary judgements should be avoided because they ignore the variability and complexity of relevance and distort the continuous nature of relevance judgements.

Borlund and Ingwersen (1998; Borlund, 2000), therefore propose the novel measures of relative recall (RR), ranked half-life (RHL), and Järvelin and Kekäläinen (2000; 2002) the measures of (discounted) cumulated gain ((D)CG), as measures augmenting the traditional ones (see also Voorhees, 2001a). An alternative to avoiding recall and precision is to generalise them so that continuous relevance judgements can be incorporated into their computation – as proposed by Kekäläinen and Järvelin (2002).

The proposed measures indeed allow for subjective non-binary multidimensional relevance assessments and may therefore augment recall and precision in evaluation. The new measures, however, like recall and precision, are just measuring devices based on any kind of relevance assessments they are supplied with. They are all immune to the way of assessing relevance (whether binary, dynamic and subjective or not). For example, Borlund (2000) states that the RHL and RR measures may well be used in non-interactive experiments within the Laboratory Model. Järvelin and Kekäläinen's (2000; 2002) measures are directly used within the Laboratory Model.

Borlund (2000) notes that there is a need for a measure bridging between the subjective assessments and objective system performance. She also correctly states that the RR measure does this (save for identified problems with differing scales of evaluation in some cases). However, traditional PR-curves already do this between any chosen kinds of subjective assessments and the objective system performance, indicated by retrieval scores of documents and their ranking (or algorithmic relevance).

As the new measures are well defined, they are usable in the evaluation of IR algorithms but they do not directly affect the design of IR algorithms.

4.3. Two Evaluation Scenarios

We shall now present two evaluation scenarios. The first one has a broader approach to IR than the Laboratory Model, paying attention to various IR strategies and tactics, and interfaces supporting them. We call it the IR Interfaces Scenario. The second is still broader, allowing for real users engaged in various types of tasks in various types of situations. We call it the Task IR Scenario. The components of the scenarios are illustrated in Figure 4.

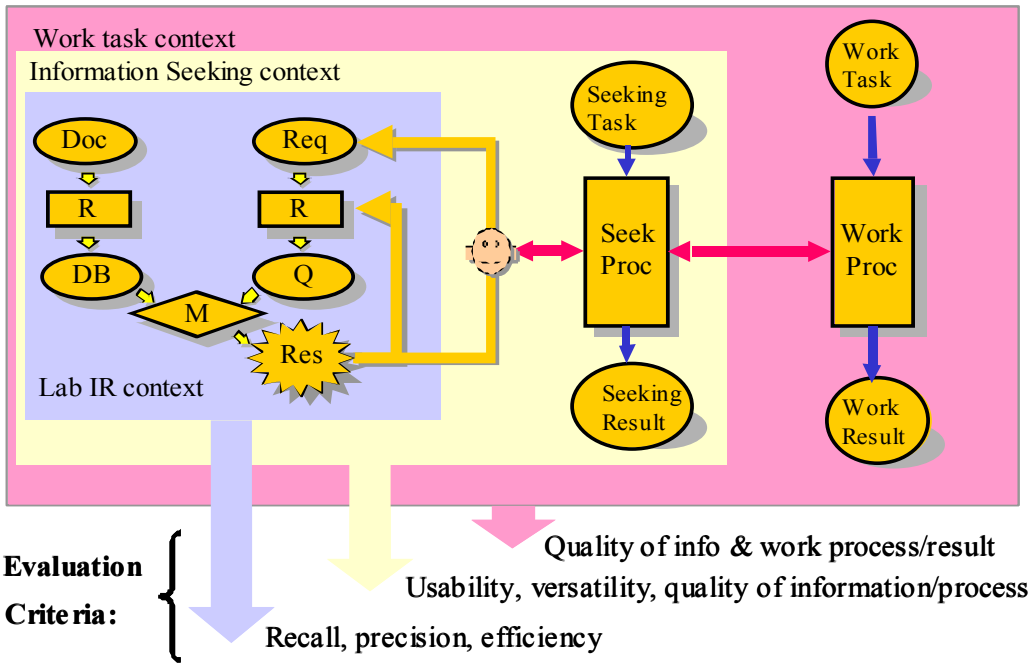

Figure 4. IR evaluation frameworks

A System Developer's View – the IR Interfaces Scenario

In this scenario, the system consists of the database(s), algorithms, requests, and users (two inner frameworks in Figure 4). Stored relevance assessments are not banned but are not the key of evaluation, either. The users are real users experienced with the kind of real or simulated tasks giving rise to search tasks. Only topical relevance is used for retrieval, unless theory and representations are developed for higher-order relevance. The scenario deliberately allows for fuzzy requests. The goal is to develop IR algorithms to (1) support the exploration of information spaces (documents, concepts) to aid request formulation, (2) identify and rank a number of relevant documents for sequences of varying requests, and (3) present such documents so as to support their judgment in terms of higher-order relevance. The success measures include criteria for usability and versatility, as well as precision and recall. The plain retrieval component may be evaluated through recall and precision, but the whole algo-

rithmic component needs broader measures. Simulated work tasks may be used as proposed by Borlund (2000).

The rationale of the framework is as follows: After all, the heart of IR is aiding the user to formulate requests, acquiring for the algorithms as much evidence as possible on relevance features, and then matching explicit request and document representations. Formally matching features suggest topical relevance, which predicts higher-order relevance (Gluck, 1996). The framework focuses on interfaces that help query formulation or allow alternative approaches. Complex dynamic interaction is a sequence of topical interactions. Recall and precision describe the filtering capability of the IR system at any stage in the process but play a lesser role (especially the former). Usability and versatility criteria should be explicated for classes of users/tasks. This is, in effect, building and testing whole systems for real users with real search tasks, and simulated work tasks. There is no guarantee however of solving any real life IR problems of user groups unless informed by studies on information seeking and retrieval of such groups.

There is empirical work (e.g., Ellis, 1989 Ellis & Haugan, 1997) suggesting that various tactics would be desirable for some user communities. On the theoretical side, e.g., Bates (1989; 1990), points out possibilities as calls for work in the area. Some of the issues may seem, from the engineering point of view, simple and thus uninteresting to computer scientists (e.g., providing for tactics like author or journal name search, or citation linking), not touching the "tough issues" of representation and matching.

The framework accounts for part of the criticism leveled against the Laboratory Model. It involves users and tasks, at least simulated tasks. A basic feature of the framework is the active user learning and assessing relevance. Recall and precision are used for the algorithmic component in isolation, other measures are looked for assessing usability and versatility.

Developing IR, not just algorithms – the Task IR Framework

In this scenario, the system consists of the database(s), algorithms, requests, users, tasks, and task processes (see Figure 4, all the frameworks). Stored relevance assessments are not banned but the focus is at higher-order relevance. The users are real users experienced with real or simulated tasks giving rise to search tasks. Only topical relevance is used for retrieval, unless theory and representations are developed. The scenario deliberately allows for fuzzy requests. The goal is to support task process, improve the quality of task outcomes through

improved IR algorithms. The success measures are case-based, usability and versatility criteria included. Recall and precision are not irrelevant but not primary, either. The plain retrieval component may be evaluated through recall and precision, but the whole algorithmic component needs broader measures. Simulated work tasks may be used as proposed by Borlund (2000).

The rationale is that, after all, IR is a support activity. By studying the information interaction of people engaged in real life tasks it is possible to learn, which aspects of interaction are amenable to IR type of processing. Database contents, representations, matching and interfaces all are tailored to the task setting. Working practices of actors are adapted to the system setting (through learning by doing or training). The algorithmic retrieval components perform as much (or as little) as can explicitly be specified and the intelligent human actors do the rest. IR may happen without explicit activation. IR becomes one function in a composite systems environment. Documents found may still be only topically relevant, if better explicit representations cannot be specified in the environment. This is, in effect, building and testing whole systems for real users with real tasks. There is a guarantee of solving typical real life problems of the user group, if anything.

It is clear that the criticisms leveled at the Laboratory Model are all accounted for. There are new ones, e.g., the findings may not be generalizable to other circumstances.

5. DISCUSSION

Higher-Order Relevance in the Evaluation of IR Algorithms?

Higher-order relevance concepts clearly play a role in IR interaction. Therefore they should be accounted for in IR evaluation. However, using them for the evaluation of IR algorithms presents difficulties, at least for the moment.

If we try to find out, whether IR algorithm A is better than algorithm B, based on situational relevance, we face some difficulties. Best-match IR systems have been designed on the basis of topical relevance solely – while some more traditional Boolean systems allow the use of bibliographic data which may relate to higher-order relevance criteria. However, there is no established theoretical connection between higher-order relevance criteria and explicit document features. This is not to say that such connections cannot exist. People use explicit situational criteria based on bibliographic attributes and citation links in relevance assessments.

Yet serving these criteria by system features is a challenge because of lack of understanding on the variability of the criteria and on the combination of evidence on the criteria. There neither is much sense in evaluating the current algorithms' performance on criteria, which they completely ignore (save for correlation of topical and higher-order relevance

If the proposals to use higher-order relevance, based on textual features, for evaluating IR algorithms were successful to show differences (after normalizing for topical relevance), it would be a real break-through. However, this would require that IR software should be able to recognize indications of higher-order relevance in document texts. Because software deals with explicit text (representations) at the level of character strings, the software should be equipped with the capability to recognize character strings in text that indicate higher-order relevance. As early as 1971 Cooper stated: "Logical relevance is almost the only factor in utility which the [system] designer does know how to deal with effectively at present. this suggestion, if true, would help to explain why topic relevance, and not utility, has received the most attention in the literature." Nothing has changed in thirty years?

Even if there were in document representations explicit features that indicate higher-order relevance, or they could be added to them, this might have just minor effects on present-day IR algorithms. The current representation and matching approaches might serve these features well. If the bibliographic and citation attributes of documents cannot be used as relevance indicators generally, the system *interfaces* might be developed to allow easy situational relevance input by the user on such attributes. The algorithms would be affected by the need to combine relevance evidence based on textual features and bibliographic and citation attributes for fair ranking. Evidence based on textual features is computed in current IR practice from textual feature occurrence statistics. Evidence based on bibliographic and citation attributes is of different nature (e.g. weighted preference lists) and has different statistical properties (citation statistics).

Interaction – Sequences of Batch Searches?

Interaction with dynamic needs can be seen as a sequence of retrieval actions where for each step there is a short-term request that can be learned. The IR algorithm should perform, as well as possible, to serve that request, taken topically, as long as features predicting anything of higher-order are beyond reach. And then serve the next, perhaps highly modified, request.

Although a user is likely to modify his/her request and relevance criteria in subsequent interactions, it has not been shown that this should affect the design of the retrieval algorithms.

Proper Systems, Proper Measures

We considered above the Laboratory Model and two evaluation scenarios, which see the system to be evaluated differently. Care should be taken in deciding, which system to evaluate and how. In the Laboratory Model, focusing on representation and matching, recall and precision seem suitable and sufficient measures (Robertson, 2000). However, if there is interaction involved, the system contains the user, and recall and precision lose their dominance. Measures for usability and learning are needed. If tasks / situations are involved, also evaluation needs a broader scope to reveal its support to task performance. Within the two scenarios we need to work toward establishing relevant evaluation criteria above the representation/matching level. Where should IR stop and Sociology start? It's a question of how far the IR community wants to go. We should not discontinue developing IR algorithms along the Laboratory Model. Other things we should consider as well are suggested in Figure 4.

Some Research Problems Requiring Attention

We finally consider research problems requiring attention, if further advances are required in realistic user/task-oriented IR evaluation, especially if this is to affect the design of the algorithmic components:

- *Relevance and representation.* We should investigate the relationship of topical relevance to the higher-order relevance. What remains there unexplained by topical relevance? How could it be (effectively and efficiently) accounted for, if at all?
- *IR interfaces.* Interface capabilities that allow users to express conditions on non-topical attributes as evidence for higher-order relevance need to be developed. Such interface capabilities should also handle the great variability of representation of, e.g., person names or citations.
- *Role of IR, goals for IR systems in various situations.* How could users be supported in learning and expressing their needs/requests, or better performing their tasks through IR methods? What kinds of interface components are required? How are the retrieval results used? How to embed IR into the users' systems environment?

- *Evaluation measures.* Above the traditional and proposed evaluation measures, how do we evaluate system effectiveness in supporting users in learning, exploring and expressing their needs/requests, or their contribution to task performance?

6. CONCLUSION

We propose as a conclusion, that the Laboratory Model is, while clearly limited in scope, not as notorious as criticized when used as a model for developing and evaluating IR algorithms for Interactive IR systems. The critics of the model are right in their claims but this rather suggests additional, broader additional evaluation scenarios than discarding the old one. Also the innovations may touch other areas than text matching between topical requests and document texts: other strategies may be relevant (or rehabilitated), system interfaces may need novel (or rehabilitated) components. Broader evaluation scenarios and further novel evaluation measures are needed. In particular, topical relevance judgment seems well founded for developing IR algorithms for representation and matching. In order to apply the higher-order relevance within the IR algorithms, advances regarding the explicit document / request features or attributes (suggesting such relevance) and their representation are required.

REFERENCES

Barry, C.L. (1994). User-defined relevance criteria: An exploratory study. *Journal of the American Society for Information Science, 45*(3), 149-159.

Barry, C.L. & Schamber, L. (1998). Users' criteria for relevance evaluation: a cross-situational comparison. *Information Processing & Management, 34*(2/3), 219-236.

Bateman, J. (1998). Changes in Relevance Criteria: A Longitudinal Study. *In Proceedings of the 61st American Society for Information Science Annual Meeting 35* (pp. 23-32).

Bates, M. (1989). The design of browsing and berrypicking techniques for the online search interface. *Online Review, 13*, 407-424.

Bates, M. (1990). Where should the person stop and the information search interface start? *Information Processing and Management*, *26*(5), 575-591.

Beaulieu, M., Robertson, S. & Rasmussen, E. (1996). Evaluating interactive systems in TREC. *Journal of the American Society for Information Science, 47*(1), 85-94.

Belkin, N. (1993). Interaction with texts: Information retrieval as information seeking behavior. In *Information Retrieval '93*. Konstanz: Universitetsverlag Konstanz.

Borlund, P. (2000). Experimental components for the evaluation of interactive information retrieval systems. *Journal of Documentation, 56*(1), 71-90.

Borlund, P. & Ingwersen, P. (1998). Measures of relative relevance and ranked half-life: Performance indicators for interactive IR. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 324–331). New York: ACM.

Burgin, R. (1992). Variations in relevance judgements and the evaluation of retrieval performance. *Information Processing & Management, 28*(5), 619–627.

Byström, K. & Järvelin K. (1995) Task Complexity Affects Information Seeking and Use. *Information Processing & Management, 31*(2), 191-213.

Cleverdon, C.W. (1967). The Cranfield tests on index language devices. *Aslib Proceedings*, *19*, 173–194.

Cosijn, E. & Ingwersen, P. (2000). Dimensions of relevance. *Information Processing & Management*, *36,* 533–550.

Cuadra, C.A. & Katter, R.V. (1967*). Experimental studies of relevance judgments: Final report. Vol. I: Project summary.* Santa Monica, CA: System Development Corporation.

Ellis, D. & Haugan, M. (1997), Modeling the information seeking patterns of engineers and research scientists in an industrial environment. *Journal of Documentation, 53*(4), 384-403.

Ellis, D. (1989). A behavioural approach to information retrieval design. *Journal of Documentation, 45*(3), 171-212.

Froehlich, T. J. (1994). Relevance reconsidered – towards an agenda for the 21st century: Introduction to special topic issue on relevance research. *Journal of the American Society for Information Science, 45*(3), 124–134.

Gaizauskas, R. & Wilks, Y. (1998). Information extraction: Beyond Document Retrieval. *Journal of Documentation*, *54*(1), 70-105.

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

Harter, S.P. & Hert, C.A. (1997). Evaluation of information retrieval systems: Approachs, issues, and methods. In *Annual Review of Information Science and Technology*, vol. 32. (pp. 3-94).

Hawking, D., Voorhees, E., Craswell, N. & Bailey, P. (2000). Overview of TREC-8 Web Track. In *Proceedings of the eight Text REtrieval Conference.* NIST Special Publication 500-246 (pp. 131-150).

Hersh, W. & Over, P. (2000). SIGIR workshop on interactive retrieval at TREC and beyond. *SIGIR Forum 34* (1). Retrieved April 25, 2002, from: http://www.acm.org/sigir/forum/S2000/Interactive_report.pdf

Hull, D. (1996). Stemming Algorithms: a case study for detailed evaluation. *Journal of the American Society for Information Science*, *47*, 70-84.

Ingwersen, P. (1996). Cognitive Perspectives of IR Interaction: Elements of a Cognitive IR Theory. *Journal of Documentation*, *51*(1), 3-50.

Ingwersen, P. & Willett, P. (1995). An introduction to algorithmic and cognitive approaches for information retrieval. *Libri, 45*, 160–177.

Järvelin, K. & Kekäläinen, J. (2000). IR evaluation methods for highly relevant documents. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 41–48). New York: ACM.

Järvelin, K. & Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. Submitted to *ACM Transactions on Information Systems*.

Kekäläinen, J. & Järvelin, K. (2002). Using graded relevance assessments in IR evaluation. To appear in *Journal of the American Society for Information Science and Technology*.

Kuhlthau, C. (1993). *Seeking Meaning.* Norwood, NJ:Ablex,

Over, P. (1997). TREC-5 Interactive Track Report. In *Proceedings of the Fifth Text REtrieval Conference*. NIST Special Publication 500-238 (pp. 29-56).

Rees, A.M. & Schultz, D.G. (1967). *A field experimental approach to the study of relevance assessments in relation to document searching.* Cleveland: Case Western Reserve University.

Robertson, S.E. (1977). The probability ranking principle in IR. *Journal of Documentation, 33*(4), 294–304.

Robertson, S.E. (2000). Salton Award Lecture: On theoretical argument in information retrieval. *SIGIR Forum, 34* (1). Retrieved January 29, 2001, from: http://www.acm.org/sigir/forum/F2000-TOC.html.

Saracevic, T. (1975). Relevance: A review of and framework for the thinking on the notion in information science. *Journal of the American Society for Information Science 26*(6), 321–343.

Saracevic, T. (1996). Relevance reconsidered '96. In *Proceedings of the Second International Conference on Conceptions of Library and Information Science: Integration in Perspective* (pp. 201–218). Copenhagen: The Royal School of Librarianship.

Saracevic, T. (1997). The stratified model of information retrieval interaction: Extension and applications. In *ASIS '97: Proceedings of the 60th ASIS annual meeting* (pp. 313–327). Medford, NJ: Information Today.

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

Schamber, L., Eisenberg, M. B. & Nilan, M. S. (1990). A re-examination of relevance: toward a dynamic, situational definition. *Information Processing & Management*, *26*(6), 755–776.

Soergel, D. (1994). Indexing and Retrieval Performance: The Logical Evidence. *Journal of the American Society for Information Science, 45*(8): 589-599.

Sparck Jones, K. (1995). Reflection on TREC. *Information Processing & Management, 31*(3), 291–314.

Tang, R., Shaw, W.M. & Vevea, J.L. (1999). Towrds the identification of the optimal number of relevance categories. *Journal of the American Society for Information Science, 50*(3), 254-264.

Vakkari, P. & Hakala, N. (2000). Changes in Relevance Criteria and Problem Stages in Task Performance. *Journal of Documentation, 56*(5), 540-562.

Vakkari, P. (2001). A theory of the task-based information retrieval process: a summary and generalization of a longitudinal study. *Journal of Documentation, 57*(1), 44-60.

van Rijsbergen, C.J. & Lalmas, M. (1996). Information calculus for information retrieval. *Journal of the American Society for Information Science, 47*, 385-398.

Wang, P. & Soergel, D. (1998). A Cognitive Model of Document Use during a Research Project. Study I. Document Selection. *Journal of the American Society for Information Science, 49*(2), 115-133.

Wilson, P. (1973). Situational Relevance. *Information Storage and Retrieval, 9*(8), 457-471.

Wilson, T. (1999). Models in information behaviour research. *Journal of Documentation*, *55*(3), 249-270.

Voorhees, E. (2001a). Evaluation by highly relevant documents. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 74 – 82). New York: ACM.

Voorhees, E. (2001b). Overview of the TREC 2001 question answering track. In *The Tenth Text REtrieval Conference (TREC 2001)*. Retrieved April 25, 2002, from: http://trec.nist.gov/pubs/trec10/papers/qa10.pdf.

Voorhees, E. & Harman, D. (2001). Overview of TREC 2001. In *The Tenth Text REtrieval Conference (TREC 2001)*. Retrieved April 25, 2002, from: http://trec.nist.gov/pubs/trec10/papers/overview_10.pdf..