Information Research, Vol. 8 No. 3, April 2003

The IIR evaluation model: a framework for evaluation of interactive information retrieval systems

Pia Borlund
Department of Information Studies
Royal School of Library and Information Science, Aalborg branch
Aalborg, Denmark

Abstract

An alternative approach to evaluation of interactive information retrieval (IIR) systems, referred to as the IIR evaluation model, is proposed. The model provides a framework for the collection and analysis of IR interaction data. The aim of the model is two-fold: 1) to facilitate the evaluation of IIR systems as realistically as possible with reference to actual information searching and retrieval processes, though still in a relatively controlled evaluation environment; and 2) to calculate the IIR system performance taking into account the non-binary nature of the assigned relevance assessments. The IIR evaluation model is presented as an alternative to the system-driven Cranfield model (Cleverdon, Mills & Keen, 1966; Cleverdon & Keen, 1966) which still is the dominant approach to the evaluation of IR and IIR systems. Key elements of the IIR evaluation model are the use of realistic scenarios, known as simulated work task situations, and the (call for) alternative performance measures. A simulated work task situation, which is a short 'cover story', serves two main functions: 1) it triggers and develops a simulated information need by allowing for user interpretations of the situation, leading to cognitively individual information need interpretations as in real life; and 2) it is the platform against which situational relevance is judged. Further, by being the same for all test persons experimental control is provided. Hence, the concept of a simulated work task situation ensures the experiment both realism and control. Guidelines and recommendations for the application of simulated work task situations are provided. Examples of alternative performance measures are: relative relevance (RR), ranked half-life (RHL) (Borlund & Ingwersen, 1998), cumulated gain (CG) and cumulated gain with discount (DCG) (Järvelin & Kekäläinen, 2000). These measures can incorporate non-binary relevance assessments, necessary due to the result of realistic interaction and relevance assessment behaviour of users in the process of searching and assessing relevance of retrieved information objects.

Introduction

Various researchers (e.g., Saracevic, 1995; Harter, 1996; Beaulieu, Robertson & Rasmussen, 1996; Ellis, 1996b; Borlund & Ingwersen, 1997; Kekäläinen & Järvelin, 2002) have expressed a demand for alternative approaches to the performance evaluation of interactive information retrieval systems (IIR systems). That is, alternative to the experimental Cranfield model which still is the dominant evaluation approach to the evaluation of IR and IIR systems. The Cranfield model derives directly from Cranfield II (Cleverdon, Mills & Keen, 1966; Cleverdon & Keen, 1966) and is based on the principle of test collections, that is: a collection of documents; a collection of queries; and a collection of relevance assessments. The Cranfield model includes also the measurement of recall and precision ratios as indicators of system performance. The Cranfield model constitutes the empirical research tradition of the development and testing of IR systems employed by the system-driven approach to IR (e.g., Swanson, 1986; Ellis, 1996a). The emphasis in this research tradition is on controlled laboratory tests. The objective of the Cranfield model is to keep all variables controlled and to obtain results, about which one can state conclusions about retrieval systems in general (Robertson 1981: 12). However, the Cranfield model suffers from limitation due to its restricted assumptions on the cognitive and behavioural features of the environment in which (I)IR systems function (Ellis 1996a: 20). These limitations are the reasons for the demand for alternative approaches to evaluation of IR and IIR systems. In brief, the demand is described and summarised with the *three revolutions*

put forward by Robertson and Hancock-Beaulieu (1992, pp. 458-459).

- The cognitive revolution;
- The relevance revolution; and
- The interactive revolution.

The *cognitive* and *relevance revolutions* require realism with reference to the formation of information need, and relevance assessment processes. This means that in the context of (I)IR evaluation an information need ought to be treated as a user-individual and potentially dynamic concept, and the multidimensional *and* dynamic nature of relevance should be taken into account, just as relevance should be judged against the information need situation, not the query or even request, and by the person who owns the information need. The *interactive revolution* points to the fact that IR systems have become more interactive. Due to the type of IR system, which IIR systems constitute, that is, systems where the user dynamically conducts searching tasks and correspondingly reacts to system responses over session time, the evaluation of IIR systems consequently has to include the user's interactive information searching and retrieval processes.

The three revolutions point to requirements that are not fulfilled by the system-driven IR evaluation approach based on the Cranfield model. The Cranfield model does not deal with dynamic information needs but treats information needs as a static concept entirely reflected by the search statement (query). This implies the assumption that learning and modifications by users are confined to the search statement alone. Furthermore, this model uses only binary and topical-oriented relevance. The conclusion is that the batch-driven mode of the Cranfield model is not suitable for the evaluation of IIR systems, which, if carried out as realistically as possible, requires human interaction, potentially dynamic information need interpretations, and the assignment of multidimensional and dynamic relevance.

It could be argued that the second main approach to IR systems evaluation, the user-oriented approach fulfils the requirements outlined. As opposed to the system-driven approach the user-centred approach defines the IR system much broader, viewing the seeking and retrieval processes as a whole. The main purpose of this type of evaluation is concerned with how well the user, the retrieval mechanism, and the database interact extracting information, under real-life operational conditions. In this approach the relevance judgements have to be given by the original user in relation to his or her personal information need which may change over session time. The assumption is that the relevance judgements represent the value of the information objects for a particular user at a particular point in time, hence the assessments can only be made by the user at that time. Further, relevance is judged in accordance to subjective situational relevance in a non-binary way. As such the requirement as to realism is fulfilled. But like the system-driven approach the user-oriented approach quantifies performance effectiveness as (relative) recall and precision ratios in spite of collecting non-binary-based relevance assessments (e.g., Lancaster, 1969).

Briefly summarised the qualities of the two main evaluation approaches are similar to the conflict issues between them (Robertson & Hancock-Beaulieu, 1992: 460). The two approaches represent different viewpoints which each aim at the same goal: reliability of the IR test performance results. To the system-driven approach reliability of the experimental results is earned through control over the experimental variables and the repeatability of the experiments. In contrast, the user-oriented approach puts the user in focus with reference to system development, design, and evaluation which is basically carried out according to the (potential) end-user's information use, retrieval, and searching behaviour with the objective of obtaining realistic results. To the user-oriented approach the results become reliable by being loyal to the IR and searching processes. So far the approaches have carried out satisfying jobs, however, this changes with the development of IIR systems as illustrated by the *interactive revolution*. IIR systems are by definition broader in scope than traditional IR systems. By incorporating the interface functionality as well as communicative behaviour by users IIR systems are defined just as broadly as in the user-oriented approach, and the focus of the evaluation is similarly wider than in non-interactive IR. The foci of IIR evaluation include all the user’s activities of interaction with the retrieval and feedback mechanisms as well as the retrieval outcome itself. The overall purpose of the evaluation of IIR systems is to evaluate the systems in a way which takes into account the dynamic natures of information needs and relevance as well as reflects the interactive information searching and retrieval processes. Thus, a hybrid evaluation approach is proposed – *a combination of elements* from the two main approaches, the issue of experimental control plus the user-individual and dynamic nature of information needs and relevance assessments – as a reasonable setting for an alternative evaluation approach to evaluation of IIR systems. In addition, the dominating use of the ratios of recall and precision for the measurement of effectiveness of IR performance forces us to reconsidered whether these measures are sufficient in relation to the effectiveness evaluation of IIR systems.

As such, the present paper contributes to the continuing development and refinement of evaluation approaches to the research area of IR by proposing a framework for evaluation of IIR systems and information searching behaviour – the so-called IIR evaluation model. The paper is organised according to the following main sections: first, the IIR evaluation model is presented, the three parts concerning data collection and data analysis – that is, 1) the basic components; 2) recommendations for the applications of simulated work task situation; and 3) alternative performance measures. The penultimate section looks into the provisional use of the IIR evaluation model – or parts of the model, and hereby verifying the need for a framework for evaluation of IIR systems. The final section closes the paper with summary statements.

The IIR evaluation model

The present paper describes the framework as an aggregated model for the evaluation of IIR systems and/or information seeking behaviour – including rationale and recommendations for application of the model – or parts of the model. During the ongoing process of developing the framework parts of the model have been described in previous publications (e.g., Borlund & Ingwersen, 1997; Borlund & Ingwersen, 1998; Borlund & Ingwersen, 1999; Borlund, 2000b). Basically, the IIR evaluation model consists of three parts:

> Part 1. A set of components which aims at ensuring a functional, valid, and realistic setting for the evaluation of IIR systems;
> Part 2. Empirically based recommendations for the application of the concept of a simulated work task situation [1]; and
> Part 3. Alternative performance measures capable of managing non-binary based relevance assessments.

Parts 1 and 2 concern the collection of data, whereas part 3 concerns data analysis. The three model parts are described in the following sub-sections.

Part 1: The components of the experimental setting

The aim of the proposed experimental setting is to facilitate evaluation of IIR systems in a way which is as close as possible to actual information searching and IR processes, though still in a relatively controlled evaluation environment. This can be achieved by the use of the *proposed components* of model part 1:

- The involvement of potential users as test persons;
- The application of individual *and* potentially dynamic information need interpretations; and
- The assignment of multidimensional *and* dynamic relevance assessments.

The basic idea is to test and evaluate by use of the users for whom a given system is developed, and, through the users' natural use and interaction with the system (or systems), to gain knowledge about the system(s). Thus, the three components are strongly interconnected. Because without the involvement of potential users as test persons, there would be no human interaction with the systems during system evaluation. Without human interaction, there would be no application of individual and potentially dynamic information need interpretations. And without human involvement and the application of individual and potentially dynamic information need interpretations, there would be no assignment of multidimensional *and* dynamic relevance. The three components are thus necessary in order to carry out evaluation of IIR systems in a way that is close to actual information searching and retrieval processes. The application of the components allows for the collection of both traditional *system-oriented data* (i.e., data about system performance) and *cognitive data* (i.e., data which inform about the behaviour of, and experiences obtained by, the test persons when working with the retrieved information objects and the system facilities). One may look at the results of every one of the iterations or only the final one. One may hence observe the number of iterations but also follow the development of the different types of relevance or relevance criteria used by the test persons during the experiment. This aspect of system evaluation makes it more forceful, as opposed to the traditional system-driven approach signified by the Cranfield model, since it provides vital data on hitherto unknown properties, like shifts of focus of the continuous interaction. The reason is that one may allow test persons simultaneously to provide different types of relevance for each information object assessed per iteration.

The second component, the application of individual and potentially dynamic information need interpretations, is

founded on the information need development and formation theory of the cognitive viewpoint in which an information need is seen as a user-individual and dynamic concept that develops as a consequence of a *problematic situation* (e.g., Wersig, 1971; Brookes, 1980; Belkin, 1980; Belkin, *et al.*, 1982; Ingwersen, 1992; 1996). Thus, the introduction of the concept of a simulated work task situation, which is essential to the experimental setting, because of its function to ensure the experiment both realism and control. Generally, the issues of experimental *realism* and *control* are ensured by the application of all three basic components. Specifically, the simulated work task situation is the realism and control ensuring device. As in real life, the simulated work task situation is to be seen as the cause of the 'breakdown situation' in the sense of Winograd and Flores (1986), a cognitive state which creates an information need which has to be satisfied in order for the user to be able to deal with the situation and move on. The issue of realism is also ensured by the involvement of test persons (potential users) who, based on the simulated work task situation develop individual and subjective information need interpretations. Individually, the test persons interactively search, modify and dynamically assess relevance of the retrieved information objects in relation to their perceptions of the information needs and the underlying simulated work task situation. Furthermore, the involvement of potential end-users as test persons provides for the possibility of applying real information needs [2] in the experiment. The application of real information needs serves a twofold purpose: 1) they act as the baseline (or control group in the sense of a classic experiment) against the simulated information needs, both at a specific test person level and at a more general level; and 2) they provide information about the systems' effect on real information needs. As for the issue of experimental control the application of simulated work task situations ensure control by being the same for all the test persons. Or said differently, because the involved test persons all search the same set of simulated work task situations, control is gained and the search results can be compared across the systems and/or system components as well as across the group of test persons.

The simulated work task situation

A simulated work task situation is a short 'cover story' that describes a situation that leads to an individual requiring to use an IR system. The 'cover-story' is, semantically, a rather open description of the context/scenario of a given work task situation. The concept of simulated work task situation derives from Ingwersen's cognitive communication models (e.g., Ingwersen, 1992: 135; Ingwersen, 1996: 9) and the application of the work task concept by Byström and Järvelin (1995) to information problem solving and information seeking processes. The simulated work task situation serves two main functions: 1) it triggers and develops a simulated information need by allowing for user interpretations of the situation, leading to cognitively individual information need interpretations as in real life; and 2) it is the platform against which situational relevance is judged (Borlund & Ingwersen, 1997: 227-228). With reference to the process of information need development the simulated work task situation more specifically helps to describe to the test persons:

- The source of the information need;
- The environment of the situation;
- The problem which has to be solved; and also
- Serves to make the test person understand the objective of the search (Borlund & Ingwersen, 1997: 229).

In our setting the simulated work task situation is a stable concept, i.e., the given purpose and goal of the retrieval. This makes experimental control possible by providing comparable cognitive and performance data in relation to simulated information needs for the same data collection, ideally across different IR techniques, but at least for one single technique. Figure 1 shows an example of a simulated situation/simulated work task situation.

Simulated situation:

Simulated work task situation: After your graduation you will be looking for a job in industry. You want information to help you focus your future job seeking. You know it pays to know the market. You would like to find some information about employment patterns in industry and what kind of qualifications employers will be looking for from future employees.

Indicative request: Find, for instance, something about future employment trends in industry, i.e., areas of growth and decline.

Figure 1. Example of a simulated situation/simulated work task situation (Borlund, 2000a; 2000b).

Simulated situation is the label of the overall frame of the *simulated work task situation* and the *indicative request*. The indicative request is a suggestion to the test person about what to search for. It is, however, not to be seen as an example of the underlying need of the particular simulated work task situation. Comparative analyses of test persons' employment of simulated situations with or without indicative requests have been made in order to verify possible consequences of the composition of the simulated situation in order to determine and recommend whether or not an *indicative request* biases the test persons' search formulations, and whether it should be included or not in the *simulated situation*. The test results showed that the indicative request does not bias the search formulations and search behaviour, and consequently it is optional whether one wants to make use of indicative requests or not. The tests and results have previously been published in Borlund (2000a; 2000b) and are summarised below in the section on Part 2 of the model.

A distinction is made between work task and search task. The work task initiates the search task, and the search task is carried out in order to meet the requirements of the work task. Allen (1996) also makes a dual distinction, however, both referring to search task. Allen (1996; 29) suggests that:

> ...one approach to task analysis is to determine what tasks are accomplished by users as they attempt to meet their information needs and how those tasks are performed. A second approach to task analysis identifies and analyzes the tasks associated with using information devices. (Allen, 1996: 29).

The distinction made by Allen is similar to how Wilson (1999; 840) differentiates between information seeking behaviour and information searching behaviour in his 'nested model of conceptual areas of information behaviour'. Further, the first approach outlined by Allen (1996) is illustrated with the research by Byström and Järvelin (1995) and Vakkari (1999). Byström and Järvelin (1995) investigate the work task complexity of public administration workers. So does Vakkari (1999), who reports on the effect task complexity has on the public administration workers' subsequent information seeking behaviour. The second approach pointed out by Allen concerns the analysis of search task, that is, the actions and activities necessary for the user to perform during IR in order to find relevant objects required by the work task as perceived by the user – and is often used as a user-oriented approach to the evaluation of IR systems as, for instance, in the Okapi projects (e.g., Walker, 1989; Beaulieu & Jones, 1998). Another example of search task analysis is the IR system evaluation carried out within the health care domain by Hersh *et al.*, (1996). The search task approach concerns what is commonly referred to as the user's seeking and retrieval behaviour, which again may be seen as a consequence of the type of work task identified in the first approach. Byström and Järvelin (1995: 194-195) divide work task complexity into five categories, ranging from a genuine decision task to an automatic information processing task, according to the pre-determinability of the information requirement of the given work task. Vakkari (1999: 834) points to the relationship between task complexity and structure of the information problem as crucial factors in determining task performance and concludes that:

> '...they are connected to the types of information people are looking for and using, to patterning of search strategies, and to choice of relevance criteria in tasks.'

The concept of task is also used within the research area of human-computer interaction (HCI) and because this area recently has overlapped the research area of IR as the result of the introduction of end-user-oriented and IIR systems the HCI task concept has started to appear in the IR literature. The HCI task concept is, as in the case of Reid (1999; 2000) denoted as 'task' only, and is defined similarly to that of search task (e.g., Allen, 1996; Diaper, 1989a; 1989b; 1989c; Preece *et al.*, 1994). HCI task analysis is the analysis of

> '…what the human has to do (or think he or she has to do) in order to accomplish a goal… Thus, we can define a **task**…as the activities required, used or believed to be necessary to achieve a goal using a particular device' (Preece et al., 1994: 411).

The overlapping research interest shared between HCI and IR is particularly in relation to the design of IR interfaces and the determination of the functionality and the level of cognitive load of already existing IR interfaces (e.g., Henninger, 1994; Brajnik, Mizzaro & Tasso, 1996; Beaulieu & Jones, 1998). However, since task-based interfaces,

in terms of being user and domain specific, have proven to be very effective (e.g., Rasmussen, Goodstein & Pejtersen, 1994; Fischer, 1995) the work task concept and approach is used within HCI, too. An example is the work by Vassileva (1995; 1996) on work and search task analysis of the users' job domain and their activities in this domain in order to implement a task-based interface to a hyper-media information system for hospitals. Another example is provided by Henninger (1994) who evaluates the interface of Codefinder by use of so-called 'task statements'. The task statements are categorised according to three problem solving levels: specific, directed and ill-defined. The levels correspond to the three different types of information needs empirically verified by Ingwersen (e.g., 1992: 116-118): the verificative information need; the conscious topical information need; and the muddled topical type of an information need. The specific and directed task statements employed by Henninger (1994) are similar to the 'topics' used in TREC [3] where as the ill-defined task statement shares characteristics with the introduced concept of a simulated work task situation. Henninger (1994) uses the task statements as a problem-solving approach to comparatively monitor the test persons use and interaction with the interfaces under investigation. For a brief review of the task concepts applied within the research areas of information seeking, IR and HCI the reader is directed to Hansen (1999) and Byström and Hansen (2002).

As mentioned, the introduced concept of a simulated work task situation derives particularly from Ingwersen (e.g., 1992; 1996) and Byström and Järvelin (1995). Ingwersen stresses, in relation to IIR, the importance of taking into account situational factors such as the work task the user is trying to perform, what the user knows about the domain, the system, their own cognitive environment and the conceptual aspects of searching in order to achieve successful and optimal IR. All nicely illustrated in his communication model (e.g., Ingwersen, 1992: 135; 1996: 9). Further, Ingwersen (1992: 207) describes how a work task mirrors tasks and problems in the work domain that may affect the individual's cognitive workspace and activities. It is the complexity of this work task that Byström and Järvelin (1995) analyse with reference to information problem solving and information seeking processes.

As such, the work task is acknowledged as essential and central to IR, therefore a potential useful tool for the evaluation of IIR systems. The concept of a simulated work task situation is an attempt to make the work task operable by providing the test persons with a context description of the work domain and the problem to solve which can be used in the evaluation of IIR systems – as speculated on by Wilson (1973: 461) with reference to the handling of situational relevance. The simulated work task situation then functions as the trigger of the test person's information need and the platform for relevance judgement, and possible information need refinement. This is in line with the cognitive theories of the information need formation and development (e.g., Taylor, 1968; Wersig, 1971; Belkin, 1980; Belkin, Oddy & Brooks, 1982; Ingwersen, 1992; 1996) and the multidimensional and dynamic nature of relevance (e.g., Swanson, 1977; 1986; Schamber, Eisenberg & Nilan, 1990; Harter, 1992; Kuhlthau, 1993; Park, 1993; Bruce, 1994; Robins, 1997; Bateman, 1998; Spink *et al.*, 1998; Tang & Solomon, 1998) who agree that the need formation is a *situation-driven phenomenon* and that the assessment of multidimensional and dynamic relevance of the information need is based on the underlying situation.

Part 2: Recommendations for the application of simulated work task situations

An evaluation of the applicability of simulated work task situations, reported on in detail in Borlund and Ingwersen (1999) and Borlund (2000a; 2000b), positively verified that the concept of simulated work task situations is recommendable for purposes of (I)IR systems evaluation. The main result of the evaluation is: that real information needs are substitutable with simulated information needs through the application of simulated work task situations – as such to be considered the main recommendation. The main, empirically-based, recommendations for the employment of simulated work task situations are as follows:

- To employ both simulated work task situations and real information needs within the same test;
- To tailor the simulated work task situations towards the information environment and the group of test persons;
- To employ either a combination of simulated work task situations and indicative requests (simulated situations), or simulated work task situations only; and
- To permute the order of search jobs [4].

In the remainder of this section, the recommendations are explained in detail. First, the recommendation:

- To employ both simulated work task situations and real information needs within the same test.

Rationale: The recommendation to employ both types of information needs (that is, simulated and real information

needs) is empirically supported by the inference statistical analyses (t-tests and chi-square tests) of differences in the test persons' treatment of the two types of information needs – revealing no difference, meaning the two types of information needs are alike. The result of no difference between the types of information needs gives evidence to the employment of simulated work task situations only, but at the same time it also allows for the inclusion of real information needs. Real information needs may function as the baseline against the simulated information needs, both at a specific test person level and at a more general level; and in addition they may provide information about the systems' effect on this type of information needs. Further, we recommend:

- To tailor the simulated work task situations towards the information environment and towards the group of test persons. The tailoring is to include:
  - A situation which the test persons can relate to and in which they can identify themselves;
  - A situation that the test persons find topically interesting; and
  - A situation that provides enough imaginative context in order for the test persons to be able to relate and apply the situation.

Rationale: Tailoring of simulated work task situations is important. Empirical results showed that the simulated work task situation that worked the best, that is, revealing a behavioural pattern of the test persons similar to the pattern of their treatment of real information needs, fulfilled the above mentioned characteristics. The results also showed that a less relatable situation, from the test persons' point of view, may to some extent be outweighed by a topically very interesting situation. Basically, tailoring of simulated work task situations is important in order to gain a trustworthy behaviour and IR interaction of the test persons. Thus, knowledge is required about the potential group of test persons in order to generate realistic and functional simulated work task situations. Further, evaluation of IIR systems by use of simulated work task situations does, in the case of a highly domain specific document collection, require that the test persons are topical domain experts. In the situation of evaluation by use of more general collections, as in the case of our main experiment (news data) (Borlund, 2000a; 2000b) no expert knowledge is required. However, the test persons ought always to share some common characteristics that make them homogeneous as a group, so that simulated work task situation can be tailored.

In addition, the empirical results revealed that the test persons' search behaviour is not affected by whether the test persons were given a simulated work task situation and an indicative request (simulated situation), or just a simulated work task situation (e.g., see Figure 1). This leads us to recommend:

- To evaluate either by use of a combination of simulated work task situations and indicative requests, or only simulated work task situations.

The test persons were asked in a post-search interview if it made any difference to them, whether they had had both an indicative request and a simulated work task situation or just a simulated work task situation 29% replied 'yes' – it made a difference, and 71% said 'no' – it made no difference (Borlund, 2000a; 2000b). All the 'yes' answers were in favour of the indicative requests. Interestingly, the test persons explained their 'yes' differently. A few of the test persons said that the indicative requests made it easier to generate the search formulations as they picked the search terms from the indicative requests. One test person said it was helpful because the indicative request helped him understand what was expected from him. Others simply stated they preferred having both. Finally, one of the test persons said he did not use the indicative request in relation to the query formulation, but had found it useful when scanning for relevant information. This indicates that the use of the indicative requests can be constructively applied in combination with the simulated work task situations.

In addition, a definition of the topic on search can be included in the simulated situation, as it was done in the feasibility study (Borlund & Ingwersen, 1997). Empirical results reported on by Spink, *et al.*, (1998: 118) support the application of a definition of the topic on search, e.g., in a test situation where a domain specific collection is applied by test persons with little knowledge of the topic. Spink, *et al.*, (1998: 118) explain that the more the test persons know about the information requiring problem, the better they can identify the need and formulate the requests/queries which results in focused retrieval.

Another recommendation concerns the issue of rotation of search jobs between the test persons so that no test persons treat the jobs in identical order. We recommend:

- To permute the order of search jobs between the test persons.

Rationale: Permutation of search jobs ought to be done for various reasons. Firstly, in order to neutralise any effect on the results caused by increasing familiarity with the experiment (system knowledge and topicality of data collection), as traditionally done within the user-oriented approach (Tague-Sutcliffe, 1992). Secondly, in order to neutralise any effect on the results due to the relevance assessment behaviour of the test persons as to the order of search jobs. The feasibility test (Borlund & Ingwersen, 1997) revealed a *significant pattern of behaviour* among the test persons in the way they carried out the relevance assessments of the retrieved documents. Indicative results of the main experiment (Borlund & Ingwersen, 1999; Borlund, 2000a) confirms the existence of a pattern of relevance assessment behaviour as to the order of search jobs. Thus, the simulated situations/simulated work task situations are to be presented to the test persons, one at the time, in such an order that none of the test persons get the same sequence of search jobs.

The final recommendation concerns the matter of pilot testing. The recommendation is not based on empirical evidence, but on practical experiences obtained while planning for and executing the reported tests. This experience leads us to recommend:

- To pilot test prior to actual testing.

Rationale: From our perspective is pilot testing mandatory when testing by use of test persons and simulated work task situations? Pilot testing provides for an opportunity to verify the essential and critical functionality of the simulated work task situations, and if necessary to modify the simulated work task situations towards the group of test persons with help from the pilot test persons. Further, we recommend to pilot test by use of both real and simulated information needs, as real information needs may inspire to 'realistic' simulated work task situations. Consequently, pilot testing is not only a test of the experimental setting and test procedure, but concerns also the design and modification of simulated work task situations.

With this sub-section regarding the second part of the IIR evaluation model, we close the aspects concerning data collection, and move on to the third part of the model, namely: data analysis and alternative performance measures.

Part 3: Alternative performance measures

Basically, the third part of the model is a call for alternative performance measures. Two performance measures (Borlund & Ingwersen, 1998; Borlund, 2000a) are introduced 1) the measure of Relative Relevance (RR), and 2) the indicator of Ranked Half-Life (RHL). The third part of the model is not limited to the RR and RHL measures, e.g., the novel performance measures by Järvelin and Kekäläinen (2000) are included as fine examples of alternative performance measures that meet the present need. The call for alternative performance measures is necessary because recall and precision, as the traditional IR performance measures, are not ideal measures for the evaluation of IIR systems. The primary reason is that these measures are based on the binary relationships between the number of relevant/not relevant, and retrieved/not retrieved information objects. Or as said by Spink and colleagues:

> ...the current IR evaluation measures are…not designed to assist end-users in evaluation of their information seeking behavior (and an information problem) in relation to their use of an IR system. Thus, these measures have limitations for IR system users and researchers. (Spink, *et al.*, 1998: 604).

From a system-driven perspective there exists no problem with the application of recall and precision. The problem arises when the measures are applied to non-system-driven settings, e.g., within the user-oriented approach. In the latter type of settings users are involved and with them the various subjective perceptions of what is relevant as well as how relevant. The measures of recall and precision do not distinguish between the different types of relevance used for relevance assessment. Just as they do not allow for a non-binary indication of how relevant the relevant information objects are, but allow only for a binary relevance representation. The employed types of relevance within the system-driven approach to IR evaluation are those of algorithmic relevance and intellectual topicality (Borlund & Ingwersen, 1998). However, as even more types of relevancy may be employed in settings involving users a need exists for performance measures, which are capable of handling and distinguishing between the different types of relevance in order to provide information as to what the different types of relevance signify in terms of IR. It is often the case in tests where non-binary relevance judgements are applied that two or more relevance categories are merged into the binary scale of relevant and non-relevant in order to facilitate the calculation of the precision and recall measures (e.g., Su, 1992).

According to Schamber (1994: 18) the relevance categories get merged because it is assumed that no information is

being lost in the merging process. To us, the merger is a result of lack of qualified performance measures that are capable of treating the users' non-binary and subjective relevance assessments. A consequence of this is also seen in the recent tendency to calculate precision as the mean of the relevance values, that is, in the case where the users' relevance assessments are indicated as numerical relevance values (e.g., Borlund & Ingwersen, 1998; Reid, 2000). Consequently, the measures of RR and RHL are introduced, followed by a discussion of related positional oriented performance measures for the comparison of best match retrieval.

The performance measures of RR and RHL

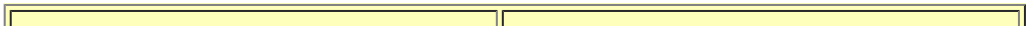
The RR measure (Borlund & Ingwersen, 1998) describes the degree of agreement between the types of relevance applied in evaluating IR systems in a non-binary assessment context. The RHL indicator (Borlund & Ingwersen, 1998), on the other hand, denotes the degree to which relevant documents are located on the top of a ranked retrieval result. The RHL performance indicator adds to the understanding of comparisons of IR best match performance by showing how well a system is capable of satisfying a user's need for information for a given set of queries at given precision levels.

The RR measure

Basically, the RR measure acknowledges the fact that different types of relevance [5] are involved in evaluation of IR systems, and especially in evaluation of IIR systems where more types of subjective relevance may be applied, and the RR measure aims at understanding this fact. The RR measure computes the degree of agreement between two results of relevance assessments (e.g., see Table 1). The two results of relevance assessments (R_1, R_2) may represents the system's output (algorithmic relevance) and the user's subjective assessments (by use of, e.g., intellectual topicality, pertinence or situational relevance) of the retrieved output. The RR measure proposes a pragmatic solution of how to bridge the gap between subjective and objective relevance – the two main classes of relevance applied to performance evaluation of IR systems, in particular IIR systems. One consequence of the multi-dimensional relevance scenario is the extent to which different types of objective and subjective relevance assessments are associated across several users and retrieval engines. Another consequence is the fact that algorithmically ranked retrieval results become interpreted and assessed by users during session time. The judgements are then in accordance with the users' dynamic and situational perceptions of a real or simulated work task situation. In addition, the assessments may incorporate non-binary relevance values. For the associative relations the suggestion is to compute a measure of relative relevance (RR) between the relevance assessments of different types of relevance by use of the cosine measure.

Hillman (1964) has suggested similar ideas of linking and describing relevance-relations by use of association measures. Also Saracevic (1984) presents similar ideas in relation to inter-search consistency. The assumption behind the proposal is that an associative inter-relationship exists between the various types of relevance which may indeed be expressed by associative relations. The RR measure is thus supposed to yield quantitative information about the performance during IIR in addition to the traditional recall and precision measures. The RR measure serves the purpose of quantifying a given relation between two types of entities, in this case between the output of two types of relevance assessments (R_1, R_2). R_1 and R_2 are constituted by the assessment values as attributed by an assessor, a user, or an engine to the retrieved information objects. Informing us about how well a system is capable of retrieving *predictable* topically relevant information objects (algorithmic relevance) and partly how well the same objects actually are subjectively relevant to the user in the sense of either intellectual topicality, pertinence or situational relevance). Further, we might learn about the nearness, as to the degrees of agreement, between the involved subjective types of relevance.

Initially, the Jaccard association coefficient was preferred to the cosine measure as the formula to use for the calculation of the RR measure (Borlund & Ingwersen, 1998). However, we find that the Jaccard measure is not capable of handling fractional counts and is abandoned for that very reason. For instance, in situations of identical match or total correspondence between the two types of relevance judgements (R_1, R_2), indicated by decimal values, the Jaccard measure does not produce a value of 1. Table 1 demonstrates a simple fictive situation of identical match between relevance assessments of R_1 and R_2, and presents the results of the Jaccard and cosine measures for the situation. Adapted and applied to the measurement of relative relevance (RR) the *Jaccard* and *cosine* formulas (e.g., Rorvig, 1999: 640) can be expressed as follows:



Jaccard: association (R_1,R_2) =	$\frac{\sum(R_1R_2)}{\sum R_1 + \sum R_2 - \sum(R_1R_2)}$
cosine: association (R_1,R_2) =	$\frac{\sum(R_1R_2)}{(\sum R_1^2)^{½} * (\sum R_2^2)^{½}}$

The value of the RR measure when calculated according to the Jaccard formula can be denoted as the intersection of the assessment values from the two types of relevance (R_1, R_2) relative to the union of the total number of assessments for a given retrieval situation. The cosine formula computes the cosine of the angle between the vector representations of the two types of relevance assessment values. The most significant difference between the two formulas is the method by which the denominator of the formulas treats the differences of relevance assessment values.

The situation illustrated in Table 1 shows a total correspondence between the relevance assessment values of R_1 and R_2. Nevertheless, the Jaccard measure fails to illustrate the perfect level of agreement as it produces an RR value of 0.619 opposite the cosine's RR value of 1. Therefore the Jaccard based RR results cannot be interpreted for the intended purpose. As shown in Table 1 it is not a problem to calculate and interpret the RR measure based on decimal values (non-binary relevance assessments) by use of the cosine formula. In the cases of total correspondence between non-binary based relevance values the cosine attains the maximum value of 1. Thus, the cosine measure is preferred to the Jaccard measure due to its robustness in a non-binary environment.

	R_1	**R_2**
doc1	0.9	0.9
doc2	0.8	0.8
doc3	0.8	0.8
doc4	0.7	0.7
doc5	0.5	0.5
sum	3.7	3.7
RR: Jaccard	0.619	
RR: cosine	1	

Table 1. Fictive data illustrating the situation of complete agreement between the non-binary relevance assessments (R_1,R_2) and the corresponding RR values of the Jaccard and cosine measures.

Consequently, the cosine measure is proposed to be used to quantify and express the degree of agreement between the two types of relevance involved (R_1 and R_2) constituted by the assessment values as attributed by an assessor (i.e., intellectual topicality), a user (e.g., situational relevance), or an engine (i.e., algorithmic relevance) to the retrieved objects. The relation between *situational relevance* and *algorithmic relevance* uncovers values which lead to: 1) an understanding of how well the perceived work task situation, assessed through situational relevance, is satisfied by the ranked output retrieved by the system; and 2) the degree to which the relevant assessments (of highly or partial relevancy) relate to the baseline measures. The lower the value, the less correspondence exists between the prediction of relevance by the system and the user's interpretation of the information objects as useful to a given task. A similar situation can be shown for the RR measure between *intellectual topicality* and *algorithmic*

relevance. We are then informed about to what extent the two types of topical related relevance assessments match each other. This tells us partly something about how well a system is capable of retrieving *predictable* topically relevant information objects, and partly how well the same objects actually are topically relevant in the intellectual sense. In cases of a high degree of equivalence in the match between the algorithmic and intellectually assessed topicality and that of algorithmic relevance and situational relevance, this fact is *no guarantee* that the aboutness of the information objects also matches the underlying purpose and work task against which the situational relevance is assessed. The relation between *intellectual topicality* and *situational relevance*, tells us about the *nearness* between the two subjective-oriented relevance types in regard to the retrieved information objects.

The RR measure generates a more comprehensive understanding of the *characteristics* of the performance of a single or several retrieval engines and algorithms in between, in particular when confronted with users. However, when comparing several systems a scaling problem exists. In relation to the application of the RR measure an issue of comparability between different engines involved in a given (I)IR experiment exists. The problem of comparison of the RR results exists due to the possible different score scales used for the indication of the assigned degrees of *algorithmic relevance* in the various engines involved in an (I)IR systems evaluation. The problem exists both in the cases of comparison of the results within the same experiment, across systems, as well as across different tests. Within the same experiment the comparability of the RR results depends on the score scale used for the relevance ranking of the algorithmic relevance. However, one may indeed compare across engines with respect to the nearness of all subjective kinds of relevance – since they are independent of the algorithmic scales used. For the comparison across tests the scale or partition of the relevance categories used for the assignment of the subjective type(s) of relevance assessments also becomes an issue. Put simply, the comparison of RR results requires that the same scale be used. Normalisation of scaling may solve this problem.

The RHL indicator

As a consequence of two or more types of relevance involved and the non-binary context in IIR the issue of *comparisons* of computed retrieval rankings become critical. By taking into account the algorithmic *rank position* and the various assigned relevance values of the retrieved information objects one takes advantage of two parameters: 1) the algorithmically ranked order which represents a list of decreasing degrees of predicted objective relevance to the user's information need; and 2) the applied subjective types and values of the relevance assessments representing the assessor's or user's interpretations of the ranked information objects. The RHL indicator makes direct use of both parameters (Borlund & Ingwersen, 1998).

The statistical method applied to calculate the Ranked Half-Life (RHL) value corresponds to the computation of the median value of grouped continuous data. The RHL value is the median “case”, i.e., the point which divides the continuous data area exactly into two parts. In nuclear physics the 'half-life' of a specific radioactive material is the time taken for half the atoms to disintegrate (Egghe & Rousseau, 1990: 267). In Bibliometrics 'cited half-life' is the time taken for half the citations to be given to a particular document (Egghe & Rousseau, 1990: 267). For the RHL indicator the time dimension is substituted by the continuous ranking of information objects produced algorithmically by a retrieval engine. Each listed information object represents a class of grouped data in which the frequency corresponds to the relevance value(s) assigned the information object.

The idea behind the application of the median point of grouped data is the fact that if top-listed information objects obtain high relevance scores assigned by the user or assessor, the median ranking for a given document cut-off and a given precision value will rise. With scattered or low placed highly relevant information objects the median 'case' will drop downwards on the original output list of objects. In the present case (Table 2), precision as performance indicator simply signifies the mean of the cumulated frequency, also used for the median calculation; but traditionally it does not inform about ranked positions. Compared to the traditional recall and precision measures the RHL indicator supplies additional information about the degree to which the engine is capable of ranking its output according to user-perceived relevance. The interpretation of RHL indicator is: the lower the RHL value, the higher on top of the rank, the better the retrieval engine for a given type of relevance.

The formula used for calculating the RHL indicator is the common formula for the median of grouped continuous data (e.g., Stephen & Hornby, 1997: 53-54):

$$M_g = L_m + \left(\frac{n/2 - \sum f2}{F(med)} \times CI \right)$$

where:

L_m = lower real limit of the median class, i.e., the lowest positioned information objects above the median class;

n = number of observations, i.e., the total frequency of the assigned relevance values;

$f2$ = cumulative frequency (relevance values) up to and including the class preceding the median class;

$F(med)$ = the frequency (relevance value) of the median class; and

CI = class interval (upper real limit minus lower real limit), commonly in IR = 1.

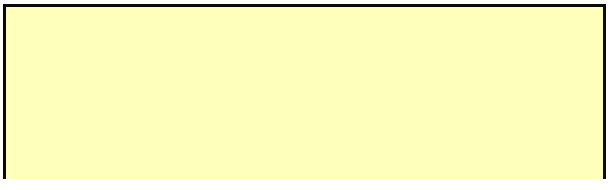
Table 2, presents RHL results computed for the purpose of illustration by use of data from the feasibility test reported on in detail in Borlund (2000a). In brief, the data used is based on the case of one test person (no. 1) and a simulated situation (a). The test person searched the Dialog Target (1993) facility (initiated by the simulated situation) and assessed the retrieval output (algorithmic relevance) according to usefulness (situational relevance). A panel [6] of two persons relevance assessed the same output in accordance to the relevance type of intellectual topicality. Further, the panel members performed Boolean Quorum searches (Lancaster & Warner, 1993) based on a direct transformation of the test person's query formulation, and assessed the outcome by intellectual topicality and situational relevance. In the present analysis the cut-off was reasonably set to fifteen documents.

a1: simulated situation a, test person no. 1

	Target					Quorum					
	algorithmic relevance	**situational relevance**	**intellectual topicality**			**situational relevance**			**intellectual topicality**		
rank order	algorithmic rank output	test person (nr. 1)	panel I	panel II	panel I+II/2	panel I	panel II	panel I+II/2	panel I	panel II	panel I+II/2
1	0.99	0.5	1.0	0.5	0.75	1.0	1.0	1.0	1.0	1.0	1.0
2	0.87	0.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.5	0.75
3	0.86	1.0	1.0	1.0	1.0	0.5	0.5	0.5	0.5	0.5	0.5
4	0.86	0.0	0.5	0.0	0.25	0.0	0.0	0.0	0.0	0.0	0.0
5	0.86	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.5	0.0	0.25
6	0.72	0.0	0.0	0.0	0.0	1.0	1.0	1.0	1.0	0.5	0.75
7	0.71	0.0	0.0	0.0	0.0	1.0	1.0	1.0	0.5	0.5	0.5
8	0.71	0.0	0.0	0.0	0.0	0.5	0.5	0.5	0.5	0.0	0.25
9	0.57	1.0	0.0	0.5	0.25	0.0	0.0	0.0	0.0	0.0	0.0
10	0.57	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
11	0.57	0.0	0.0	0.5	0.25	1.0	1.0	1.0	0.5	1.0	0.75
12	0.57	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
13	0.57	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
14	0.57	0.5	1.0	0.5	0.75	0.0	0.0	0.0	0.0	0.0	0.0
15	0.57	0.0	0.5	0.5	0.5	0.0	0.0	0.0	0.0	0.0	0.0
sum	10.575	3	5	4.5	4.75	6	6	6	5.5	4	4.75
precision	0.71	0.2	0.33	0.3	0.32	0.4	0.4	0.4	0.37	0.27	0.32
RHL-indicator	6.18	3	2.5	2.75	2.63	5.5	5.5	5.5	4.5	3	4.52
RHL index	8.7	15	7.58	9.17	8.22	13.75	13.75	13.75	12.16	11.11	14.13

Table 2. The distribution of the percentage values of the three types of relevance assessments of version a1, including the RHL values – for the Target and Quorum engines.

The Target engine achieves a precision value of 0.2 associated with situational relevance (test person no. 1) and a RHL indicator value of 3. This means that, for this test person, the Target engine is capable of providing half the cumulated relevance frequency of the 15 assessed documents within the first three listed documents. For the Quorum engine, however, the situational RHL indicator value is 5.5 for the same simulated situation as assessed by both panel members; the precision is of higher value (0.4) than for Target. Identical precision values of 0.32 with reference to intellectual topicality are attained for Target and Quorum by the panel. However, the corresponding RHL values are in favour of Target with a value of 2.63 as opposed to 4.52 for Quorum. Despite the identical precision values (0.32), which might leave us with the impression of the two engines being equally good (or bad) at retrieving topically relevant documents, it is shown that from a RHL and a user's point of view that Target is better. That is, the higher the engine can place relevant information objects the better the system. Thus, compared to ordinary precision measures the RHL indicator supplies additional valid information about the *degree* to which the engine is capable of ranking its output according to user-related relevance. As such this is a good example of how RHL supplements, in the present case precision, but potentially both precision and recall by providing additional performance information. Further it serves as a good example for the demonstration of how to calculate the RHL indicator which again shows how the indicator functions.



	intellectual topicality	
	Target	**Quorum**
rank order	panel I+II/2	panel I+II/2
1	0.75 (0.75)	1.0 (1.0)
2	1.0 **(1.75)**	0.75 (1.75)
3	**1.0** (2.75)	0.5 (2.25)
4	0.25	0.0 **(2.25)**
5	0.0	**0.25** (2.50)
6	0.0	0.75
7	0.0	0.5
8	0.0	0.25
9	0.25	0.0
19	0.0	0.0
11	0.25	0.75
12	0.0	0.0
13	0.0	0.0
14	0.75	0.0
15	0.5	0.0
sum	4.75/2=**2.38**	4.75/2=**2.38**
precision	0.32	0.32
RHL	2.63	4.52

$$\text{RHL (intellectual topicality, Target, panel)} = 2 + \left(\frac{2.38 - 1.75}{1.0} * 1\right) = 2.63$$

$$\text{RHL (intellectual topicality, Quorum, panel)} = 4 + \left(\frac{2.38 - 2.25}{0.25} * 1\right) = 4.52$$

Table 3. Extraction of Table 2 for the demonstration of calculation of the RHL indicator.

For the purpose of the demonstration an extract of Table 2 of the relevant data constitutes Table 3. The figures used for the calculation of the actual RHL values are stressed in bold, and are further shown in the formula applied. For a given document cut-off one might prefer to obtain a RHL *index* value which equals the computed RHL value normalised for the corresponding precision value (precision = 1.0). Table 2 presents the matching RHL index values. The index serves to emphasise the characteristics of the engines' ranking capabilities for the same precision values across relevance types. Thus, for a given document cut-off and a given value of precision the RHL indicator can be examined across all test persons and all test tasks for each of the involved types of relevance. Realistically speaking, in IIR experiments the document cut-offs might vary according to the engagement of each test person – a situation which then has to be normalised.

Just as the RR and RHL measures supplement the measures of recall and precision they supplement also each other. One may say that the RR measure bridges horizontally across applied types of relevance; whereas RHL indicates the vertical position of the median value of the user's assigned relevance values based on the ranked retrieval output. The RR measure as well as the RHL indicator can obviously be applied to non-interactive IR experiments like TREC, which include algorithmic rankings and assessors' relevance values assigned to these rankings. In TREC-like experiments the RR measure can be used directly to the two different types of assessments: algorithmic relevance and intellectual topicality, for the *same* retrieval engine. Across retrieval engines, the comparability of the RR results depends on the score scale used for relevance ranking of the algorithmic relevance. This can basically be explained with the different cognitive origin of the possible different retrieval algorithms. Normalisation of the scales involved may solve this problem. The RHL indicator is applicable across systems – but either directly on the

algorithmic level or limited to the subjective level alone.

Related positional oriented single performance measures

Both the RR measure and the RHL indicator represent novel approaches to the measurement of IR performance. But in contrast to the case of the RR measure which has no preceding tradition or attempts to measure IR performance across the applied types of relevance – binary or non-binary, other positional oriented IR performance measures like RHL exist. Of positional oriented single performance measures the *expected search length* (ESL) measure by Cooper (1968) is without doubt the most known. This is further indicated by the modification of the ESL by Dunlop (1997) resulting in the so-called *Expected Search Duration* (ESD). Of recent approaches, the measure of *average search length* (ASL) by Losee (e.g., 1996; 1998) will also be addressed.

The ASL measure developed by Losee (e.g., 1996; 1998) is based on the binary approach to relevant ranked information objects. At a document level

> "...the ASL represents the expected number of documents obtained in retrieving a relevant document, the mean position of a relevant document". (Losee, 1996: 96)

The ASL is somewhat similar to the ESL (Copper, 1968) and ESD (Dunlop, 1997) by calculating the mean value of the number of information objects. But they differ in the sense that the ESL and ESD are founded on the number of non-relevant information objects that the user has to view in the process of retrieving a chosen relevant information object, whereas the ASL counts the number of relevant information objects. Or said differently, the ESL measure indicates the amount of wasted search effort to be saved by using the retrieval system as opposed to searching the collection purely at random until needed relevant information objects are found (Cooper, 1968: 30); in contrast the ASL measure indicates the expected position of a relevant information object in the ranked list of information objects (Losee, 1998: 89-90). All the measures, including the RR measure and the RHL indicator, are relatively easily computed and comprehended analytically and also easily interpreted by the end user. In particular, the ESL, ASL and the RHL are related in terms of being single performance measures which work at the statistically descriptive level by treating the central tendency of retrieval success. The approaches put forward by Cooper (1968), Losee (e.g., 1996; 1998), and Dunlop (1997) all share the system-driven binary approach to the calculation of IR performance. The performance measures by Järvelin and Kekäläinen (2000) may be seen as a response to the call for alternative performance measures. Järvelin and Kekäläinen (2000: 41) introduces the three proposals as

> "…(1) a novel application of P-R-curves and average precision computations based on separate recall bases for documents of different degrees of relevance, and (2) two novel measures (CG and DCG) computing the cumulative gain the users obtain by examining the retrieval result up to a given ranked position". (Järvelin & Kekäläinen, 2000: 41)

They describe their motivation for the application of P-R-curves at individual recall levels with reference to how the traditional IR system performance evaluation is based on average precision over recall levels and P-R-curves which does not take into account the multiple degree of relevance assessments. They continue

> "...even if the original assessments may have had multiple degrees, these are generally collapsed into two for evaluation. In order to see the difference in performance between retrieval methods, their performance should be evaluated separately at each relevance level. For example, in case of a four-point assessment (say, 0 to 3 points), separate recall bases are needed for highly relevant documents (relevance level 3), fairly relevant documents (relevance level 2), and marginally relevant documents (relevance level 1). The rest of the database is considered irrelevant (relevance level 0)." (Järvelin & Kekäläinen, 2000: 42)

As such, they demonstrate that non-binary relevance is applicable within the system-driven context of IR performance evaluation.

The two measures of cumulative gain, that is, 'cumulated gain' (CG) and 'cumulated gain with discount' (DCG) seek to estimate the cumulative relevance gain the user receives by examining the retrieval result up to a given rank (Järvelin & Kekäläinen, 2000: 41). As such, both the measures are positional oriented performance measures and related to the measures of ESL, ASL and RHL. The CG and DCG are defined as follows (Järvelin & Kekäläinen, 2000: 43):

$$CG[i] = \begin{cases} G[1], \text{ if } i = 1 \\ CG[i-1] + G[i], \text{ otherwise} \end{cases}$$

$$DCG[i] = \begin{cases} G[1], \text{ if } i = 1 \\ CG[i-1] + G[i] / {}^{b}\log i, \text{ otherwise} \end{cases}$$

The CG and DCG measures differ from each other in their way of comparing the ranked output as a result of the processing of a query. Järvelin and Kekäläinen (2000: 42) explain how the CG measure builds upon the fact that highly relevant documents are more valuable than marginally relevant documents. The DCG measure takes into account the relevance assessment behaviour of users, that is, the lower the ranked position of a relevant document or partially relevant document the less valuable it is for the user, because the less likely it is that the user will examine the document.

Järvelin and Kekäläinen (2000: 43) compare the characteristics and qualities of the CG and DCG measures to those of RHL and ASL. In regard to the ASL measure (Losee, 1996; 1998) Järvelin and Kekäläinen point out how ASL is based on the concept of binary relevance, indirectly meaning that the measures of CG (and DCG) as well as RHL are 'better' as they can handle non-binary relevance assessments. The second point concerns how the CG (and DCG) has a clear advantage compared to the measures of ASL and RHL. The case is that at

> "…any number of retrieved documents examined (rank), it gives an estimate of the cumulated gain as a single measure no matter what is the recall base size." (Järvelin & Kekäläinen, 2000: 43)

The ASL measure only gives the average position of a relevant document and the RHL measures gives the median point of accumulated relevance for a given query. The third point concerns the measures' dependency or robustness with reference to outliers, i.e., relevant documents found later in the ranked order. It is explained that the measure of CG

> "…is not heavily dependent on outliers…since it focuses on the gain cumulated from the beginning of the result. The ASL and RHL are dependents on outliers although RHL is less so." (Järvelin & Kekäläinen, 2000: 43)

The fourth point stated is about the easiness and obviousness by which the CG measure can be interpreted, and further how it is more direct than P-R-curves and does not mask bad performance. No comparative comments are made to the ASL, RHL or the RR measures in this respect. However, all three measures are intuitively easy to interpret and neither of them serves the purpose of masking bad IR performance. One comment is made about the RHL, which is that it alone is not sufficient as an IR performance measure. Let us emphasise that this has never been the intention. The measures of RHL and RR are thought to supplement the measures of recall and precision and to emphasise the call for alternative performance measures, and might as well supplement the method and measures proposed by Järvelin and Kekäläinen (2000). In addition, the DCG measure is correctly said to have the following advantages not provided by the ASL or RHL measures: 1) it realistically weights down the gain received through the documents found later in the ranked list of results; and 2) it allows modelling the user persistence in examining long ranked result lists by adjusting the discount factor (Järvelin & Kekäläinen, 2000: 43).

Like the RHL indicator the proposed measures of CG and DCG require that the same category scale of subjective relevance is used when comparing across systems. The measures by Järvelin and Kekäläinen (2000) are constructive attempts of how to handle non-binary relevance assessments experienced in IR tests that involve users or test persons, e.g., tests of IIR systems. Thus, we welcome and include their proposals, that is, the novel application of P-R-curves reflecting the various degrees of relevance and the two position measures of cumulative gain of perceived relevance. All together their proposals and ours form an arsenal of alternative performance measures that fit the third part of the IIR evaluation model.

Provisional use of the IIR evaluation model

Though the proposed model to evaluation of IIR systems and user search behaviour is relatively new, parts of the

model have already been employed and reported on in the research literature (e.g., Jose, Furner & Harper, 1998; White, *et al.*, 2001; Rodden, *et al.*, 2001; Fernström & Brazil, 2001; Ruthven, 2002; Uhrskov, 2002; Nielsen, 2002; Blomgren, *et al.*, 2002). In this section we point to these IR studies and experiments, and hereby indirectly verify the need for an alternative framework for evaluation of IIR systems. It is the explicit use of simulated work task situations that is the most employed part of the model. However, it could be argued that the use of simulated work task situations to the evaluation of (I)IR systems is not new. For instance, as long ago as 1967 Cuadra and Katter (1967) reported on an IR relevance test of the variable of 'implicit use orientations' where intermediaries were given 'cover stories' of information requirements. The intermediaries used 'cover stories' for the purpose of imagining a given situation in order to help finding relevant information for the possible user in that situation. Similarly, Brajnik *et al.*, (1996) provided their test persons with descriptions of so-called 'information problems' when evaluating an IR user interface. Based on the information problem descriptions the test persons had 30 minutes to retrieve suitable documents that satisfied the described story situation. However, no methodological argumentation exists in the cases of Cuadra and Katter (1967) or Brajnik, *et al.*, (1996) of why the approaches were chosen as well as theoretical and empirical evidence to support the use of 'cover stories'.

Further, the use of simulated work task situations can be seen as related to vignettes, which are used at various stages in the data collection and analysis of information seeking behaviour. The technique is scenario based and aims at either eliciting details about information behaviour from users or displaying data meaningfully (Urquhart, 2001). Vignettes are short stories presented to the test persons, who are then asked to describe their possible reactions in response to the presented situation (Urquhart, 1999; 2001). The vignette technique shares characteristics with the psychology-founded method of empathy-based stories (e.g., Eskola, 1988) also referred to as 'non-active role-playing' (e.g., Ginsburg, 1979). This method involves the writing of short essays, according to instructions given by the researcher. The instructions are given to the test person in terms of a 'script'. So where we propose to use simulated work task situations as the trigger of simulated information needs and as the platform for the assessment of situational relevance in the context of IR, is the script used as the guide of essay writing. The script contains instructions and directions for the test person to follow and use when writing the essay. The job of the test person is either to continue the writing of a story introduced by the researcher, or to describe what has taken place prior to a given point of time in the story as outlined by the researcher. At least two different versions of the same script are employed because variation in scripts is central to the method. Within information science this approach has been employed by, e.g., Halttunen and Sormunen (2000) in the test of the computer supported learning environment called the 'IR Game'.

Jose *et al.*, (1998) are the first to evaluate by use of simulated work task situations when carrying out performance evaluation of a spatial querying-based image retrieval system. Also Reid (1999; 2000) adopts the idea of a simulated work task situation, though simply naming it 'task'. Reid (1999; 2000) is at an analytical level concerned with how to evaluate IR systems according to search task processes and subsequent information use based on simulated work task situations, which makes her propose to use a task-oriented test collection. White, *et al.*, (2001) use simulated work task situations to evaluate the WebDocSum IR system, which is a query-biased web page summariser. Comparative performance evaluation of three different versions of the WebDocSum systems is carried out also by use of simulated work task situations (White *et al.*, 2002). Their evaluation uses three different types of simulated work task situations initiating three different types of searches according to the information needs triggered, that is, fact search, decision search, and background search. The applied types of simulated work task situations correspond, to some degree, to the task categories of automatic information processing task, normal information processing task, normal decision task, and known genuine decision task presented by Byström and Järvelin (1995). At the same time the different tasks produce different types of simulated information needs within the range of the verificative and the conscious topical information needs (Ingwersen, 1992). Further, by employing different types of simulated work task situations White and colleagues (2002) expand the concept of simulated situations, that is, the use of simulated situations in engendering realistic searching behaviour for this range of tasks has not yet been tested. Ruthven (2001; 2002) employs the concept of simulated work task situations to his doctoral work when evaluating relevance feedback taking into account users' relevance assessment behaviour. Rodden *et al.*, (2001) test whether the presentation of image retrieval output arranged according to image similarity is recommendable, and hereby employ simulated work task situations. Whereas, Fernström and Brazil (2001) investigates information searching behaviour of test persons browsing sound files, initiating the browsing by use of simulated work task situations. Also Uhrskov (2002) applies simulated work task situations in her comparative study of IR searching behaviour of university students from two different science disciplines. Most recently, Nielsen (2002) has employed simulated work task situations, in the context and domain of the pharmaceutical industry, in regard to the verification of effectiveness of a thesaurus constructed by use of the word association method. As for the application of alternative performance measures, the DCG measure (Järvelin & Kekäläinen,

2000) has recently been validated by Voorhees (2001) presumably to test its fitness to TREC. Further, the INEX initiative (Fuhr et al., 2002; http://qmir.dcs.qmw.ac.uk/INEX /index.html) for the evaluation of XML retrieval, which relevance assess according to four relevance categories may consequently apply alternative non-binary based performance measures in the future computation of retrieval performance. So far the work (in progress) by Blomgren *et al.*, (2002) is the first incident of the application of the entire IIR evaluation model, that is, all three parts of the model – the components, the recommendations, and one of the alternative performance measures (RHL).

Summary statements

The IIR evaluation model is primarily a response to the domination of the Cranfield model and its application to the evaluation of IIR systems within the system-driven approach to IR evaluation. With IIR systems being defined as systems developed for and founded on *interaction*, this type of systems cannot be optimally evaluated in a static mode like that of the Cranfield model. The Cranfield model is a laboratory-based approach to systems evaluation that requires no potential users to work with the systems under testing. This again has the consequence that the Cranfield model does not deal with dynamic information needs but treats information needs as a static concept entirely reflected by the query (search statement). Furthermore, this model uses only binary topical relevance ignoring the fact that relevance is a multidimensional and potentially dynamic concept. Hence, the batch-driven mode of the Cranfield model is not suitable for the evaluation of IIR systems which, if carried out as realistically as possible, requires human interaction, potentially dynamic information need interpretations, and the assignment of multidimensional *and* dynamic relevance. These requirements are signified by the proposal of a *set of components* that constitutes the first part of the model:

- the involvement of potential users as test persons;
- the application of individual and potentially dynamic information need interpretations deriving from, e.g., the sub-component of a simulated work task situation; and
- the assignment of multidimensional *and* dynamic relevance judgements.

The set of components combined with the second part of the model, recommendations for the application of simulated work task situations, provides an experimental setting that enables to facilitate evaluation of IIR systems as realistically as possible with reference to actual information seeking and retrieval processes, though still in a relatively controlled evaluation environment. The third and final part of the model is a call for alternative performance measures that basically are capable of managing non-binary based relevance assessments, subsequent to the application of the model part no. one and two.

The actual procedure of the application of the model includes the recruitment of test persons to participate in the pilot and main experiment. To inform the test persons about when and where to show up as well as to ask them to prepare a real information need in advance of the experiment (an information need that is capable of being met by the particular collection of information objects in use, that is, if one decides to include also real information needs).

An experimental setting, in this context, necessarily includes a database of information objects as well as the system(s) under investigation. However, these components are not explicitly dealt with here. It is assumed that the system to be tested, other technical facilities, and the facilities of where to carry out the experiment are already arranged for. Knowing about the collection of information objects' topical domain (e.g., management, medicine, or art) or type of information (e.g., news, scientific, or fiction) and the characteristics in common of the group of test persons the simulated work task situations can be generated. One way to gather information and verify characteristics of the users within the particular domain is through interviews prior to generation of simulated situations, but also through pilot testing by asking the test persons to bring with them real information needs. Prior to the pilot and main experiment decisions about methods of data collection (e.g., transaction log, questionnaire, interview, observation by human or by video recording) have to be made. Decisions about the number of relevance categories to be employed have to be made as well as how to collect and capture the relevance information (e.g., electronically by log; or manually on a sheet of paper). These decisions concern the design of the experiment.

When the experiment has been designed, according to the purpose of the experiment, the experimental procedure can be verified through pilot testing, which ought to identify the requisites necessary for the experiment. The requisites may, for instance, include a dictionary, pen and paper, but also a questionnaire for the collection of demographic as well as supplementing information about the participating test persons; and a structured post-search interview to

follow-up on the test persons' experiences and perceptions of their participation in the experiment.

When the test persons show up for the experiment it is essential they get the same treatment with reference to the introduction to the experiment, the demonstration of IR system(s), and the conditions under which the experiment takes place. The experiment may start with collecting demographic information about the test persons, this may also help the test persons to ease up and become comfortable with the entire situation. A (brief) demonstration may be given of the IR system(s) and the search by the use of simulated situations/simulated work task situations can start.

The simulated situations/simulated work task situations are presented one at the time to the test persons in such an order that none of the test persons get the same sequence of search jobs. This is done to neutralise any effect on the results caused by increasing familiarity with the experiment (e.g., Tague-Sutcliffe, 1992; Borlund & Ingwersen, 1997; 1999). Further, it is required that the test persons see the same simulated situation/simulated work task situation only once, and that all test persons search all simulated situations/simulated work task situations. Just as all test persons must search all IR systems under testing (if multiple), and all simulated situations/simulated work task situations must be searched against all IR systems. If not, the results cannot be compared across the IIR systems and the group of test persons. No matter the experimental conditions and set up, that is, whether the facilities are so that several test persons can carry out searches at the same time or only one test person at the time, the test persons are not supposed to discuss the simulated work task situations with each other. The reason is to avoid the possible influence on the interpretation of the simulated work task situation. The simulated work task situations are to be perceived as unbiased as possible by the test persons. This condition entails the sharing of the overall universe in a cognitive sense by scenarios, test collections and test persons. Based on the test persons' individual perceptions of the problems described in the simulated work task situations and in accordance with the individual test persons' knowledge states they formulate information statements, which are put to the systems(s) as requests. The requests function as indications of the perceived information needs of the test persons.

According to the purpose of the experiment the test persons assess relevance of the retrieval output in conformity with the purpose of the experiment. That is, if the purpose is to compare IR systems performance or to let the test persons focus on specific actions then an 'information object cut-off' can be incorporated into the design of the experiment. If the purpose, in contrary, is to gain knowledge about the test persons' searching and IR behaviour in the process of satisfying their (simulated) information needs by use of the particular systems under investigations then no 'information object cut-off' ought be incorporated, but instead it should be decided whether or not time constraints should be built into the experimental design. The possibility of obtaining additional information about the test persons' perception of the system(s), their perceived satisfaction of their information needs, and their perceived realism of the simulated work task situations is available after the completion of each of the search jobs prior to the next search job. This can be done as a structured post-search interview or as a system built in questionnaire, which in both cases the test persons have to complete before continuing the experiment. Just as an overall post-search interview can be employed to unite and close the experiment for each participating test person.

The experimental procedure of testing can be divided into two steps. The first step concerns the collection of the experimental data, the second the analysis of the collected data – including the calculation of IR system performance. Due to the experimental components, the human participation, the dynamic information need interpretations, and the multidimensional *and* dynamic relevance assessments the relevance outcome, which provide the foundation for the calculation of the IR system performance, is different as opposed to the relevance outcome of traditional system-driven IR experiments. Traditionally, recall and precision ratios are calculated as the indicators of the IR performance. These ratios are founded on the binary relationships between the number of objectively relevant/not relevant, and retrieved/not retrieved information objects. The relevance outcome obtained by use of the proposed experimental components can make use of the measures of relative recall and precision. However, relative recall and precision can be calculated only if the relevance outcome of two or more relevance categories gets merged with the consequence of loosing the information about how relevant the retrieved and relevant information objects really are. Furthermore, the relevance outcome of the present proposed setting represents different types of relevance, as a minimum one objective and one subjective type of relevance – just as in traditional system-driven IR experiments. These different types of relevance are traditionally not distinguished between but simply treated as the one and same type. This is a problem because the different types of relevance represent *different degrees* of intellectual interpretations containing information about the IR system's capability of satisfying different types of information problems at different stages of the information seeking process (Kuhlthau, 1993). The conclusion is that IR performance based on the relevance outcome obtained as a result of the proposed experimental setting, which reflect the dynamic information searching and IR processes necessary for IIR system evaluation, either cannot be calculated or are difficult to calculate by use of the traditional static and two-dimensional performance measures of

recall and precision. Thus, we bring attention to the need of alternative and complementary performance measures to complete the IIR evaluation model. The measures of relative relevance (RR), ranked half-life (RHL) (Borlund & Ingwersen, 1998; Borlund, 2000a), cumulated gain (CG) and cumulated gain with discount (DCG) (Järvelin & Kekäläinen, 2000) are examples of such alternative measures which can handle and compare different types of relevance and manage non-binary relevance assessments.

In the presentation of the IIR evaluation model two issues have been emphasised: 1) the essentiality of the sub-component of the simulated work task situation to the experimental setting because of its function to ensure the experiment both realism and control; and 2) how the employment of the proposed components changes the experimental relevance outcome making recall and precision insufficient for the measurement of IIR performance.

We see the proposed evaluation model as a first instance of a cognitive approach to the evaluation of IIR systems. The model is anchored in the holistic nature of the cognitive viewpoint by being a hybrid of the two main approaches to IR systems evaluation – the system-driven and the cognitive user-oriented approaches – building upon each their central characteristics of control *and* realism.

Notes

1. With respect to the traditionally employed performance measures of recall and precision.
2. A real information need is defined by the user and it is characterised by being of personal interest and importance to the user.
3. TREC is the acronym for Text REtrieval Conferences. For an overview of TREC the reader is directed to the TREC homepage at: http://trec.nist.gov/pubs.html, where overview papers can be found.
4. 'Search job' is used as a common expression of the simulated work task situations and the test persons' real information needs.
5. For definitions of different types of relevance see, e.g., Saracevic (1996); Borlund & Ingwersen (1998); Cosijn & Ingwersen (2000).
6. A panel is a group of domain experts or members of the research team, but as opposed to a control group it is the panel that has generated the simulated situations/simulated work task situations in advance.

Acknowledgement

The author thanks lecturer Dr. Ian Ruthven and doctoral student Jesper W. Schneider for their careful reading and constructive comments to an earlier draft of the paper. The author is also grateful to Professor Peter Ingwersen for many fruitful discussions.

The present work is carried out as part of the TAPIR research project headed by Professor Peter Ingwersen at the Royal School of Library and Information Science, Department of Information Studies. The work is financially supported by a grant from the Research Fund of the Danish Ministry of Culture (ref. no.: A 2001 06-021(a)).

References

- Allen, B.L. (1996) *Information tasks: towards a user-centered approach to information systems*. San Diego, CA: Academic Press.
- Beaulieu, M. & Jones, S. (1998) Interactive searching and interface issues in the Okapi best match probabilistic retrieval system. *Interacting with Computers*, **10** (3), 237-248.
- 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.J. (1980) Anomalous states of knowledge as a basis for information retrieval. *Canadian Journal of Information Science*, **5**, 133-143.
- Belkin, N.J., Oddy, R. & Brooks, H. (1982) ASK for information retrieval: part I. Background and theory. *Journal of Documentation*, **38**(2), 61-71.
- Blomgren, L., Vallo, H. & Byström, K. (2002) Evaluation of an information system in an information seeking process: the preliminary results. (Unpublished: work in progress). http://user.tninet.se/~zpd318f/paper.htm
- Borlund: (2000a) *Evaluation of interactive information retrieval systems* Doctoral dissertation: Åbo Akademi University. (Åbo (Turku): Åbo Akademi University Press)

- Borlund, P. (2000b) Experimental components for the evaluation of interactive information retrieval systems. *Journal of Documentation*, **56**(1), 71-90.
- Borlund, P. & Ingwersen, P. (1997) The development of a method for the evaluation of interactive information retrieval systems. *Journal of Documentation*, **53**(3), 225-250.
- Borlund, P. & Ingwersen, P. (1998) Measures of relative relevance and ranked half-life: performance indicators for interactive IR. In: Croft, B.W, Moffat, A., van Rijsbergen, C.J., Wilkinson, R., and Zobel, J., eds.
- Brajnik, G., Mizzaro, S., & Tasso, C. (1996) Evaluating user interfaces to information retrieval systems: a case study on user support. In: Frei, H.P., Harman, D., Schäuble, P. & Wilkinson, R., eds. *Proceedings of the 19th ACM Sigir Conference on Research and Development of Information Retrieval, Zurich, 1996.* pp. 128-136. Konstanz: Hartung-Gorre.
- Brookes, B.C. (1980) The foundation of information science: part I: philosophical aspects. *Journal of Information Science*, **2**, 125-133.
- Bruce, H.W. (1994) A cognitive view of the situational dynamism of user-centered relevance estimation. *Journal of the American Society for Information Science*, **45**(3), 142-148.
- Byström, K. & Hansen, P. (2002) Work task as units for analysis in information seeking and retrieval studies. In: Bruce, H., Fidel, R., Ingwersen, P. & Vakkari, P., eds. *Emerging Frameworks and Methods, Seattle, 2002.*. pp. 239-251. Colarado: Libraries Unlimited.
- Byström, K. & Järvelin, K. (1995) Task complexity affects information seeking and use. *Information Processing & Management*, **31**(2), 191-213.
- 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.
- Cosijn, E. & Ingwersen, P. (2000) Dimensions of relevance. *Information Processing & Management*, **36**(4) 533-550.
- Cuadra, C.A. & Katter, R.V. (1967) Opening the black box of 'relevance'. *Journal of Documentation*, **23**(4), 291-303.
- Diaper, D. (1989a) (Editor) *Task analysis for human-computer interaction.* Chichester: Ellis Horwood.
- Diaper, D. (1989b) Task analysis for knowledge descriptions (TAKD): the method and an example. In: Diaper, D., ed. *Task analysis for human-computer interaction.* pp. 108-159. Chichester: Ellis Horwood,
- Diaper, D. (1989c) Task observation for human computer interaction. In: Diaper, D., ed. *Task analysis for human-computer interaction* . pp. 210-237. Chichester: Ellis Horwood,
- Dunlop. M. (1997) Time, relevance and interaction modelling for information retrieval. In: Belkin, N.J., Rarasimhalu, A.D., & Willett, P., eds. *Proceedings of the 20th ACM SIGir Conference on Research and Development of Information Retrieval*. Philadelphia, 1997. pp. 206-213. New York, N.Y.: ACM Press,
- Egghe, L. & Rousseau, R. (1990) *Introduction to informetrics: quantitative methods in library and information science. /i> Amsterdam: Elsevier Science Publishers.*
- *Ellis, D. (1996a) Progress and problems in information retrieval. London: Library Association Publishing.*
- *Ellis, D. (1996b) The dilemma of measurement in information retrieval research. Journal of Documentation, **45**(3), 23-36.*
- *Eskola, A. (1988) Blind alleys in social psychology. Amsterdam: Elsevier Science Publishers.*
- *Fischer, G. (1994) New perspectives on working, learning, and collaborating and computational artefacts in their support. In: Böcker, H.D., ed. Proceedings Software-Ergonomie '95. pp. 21-41. Stuttgart: B.G. Teubner Verlag,*
- *Fuhr, N., Gövert, N., Kazai, G. & Lalmas, M. (2002) INEX: initiative for the evaluation of XML retrieval. http://ls6-www.informatik.uni-dortmund.de/bib/fulltext/ir/Fuhr_etal:02a.pdf*
- *Halttunen, K. & Sormunen, E. (2000) Learning information retrieval through an educational game: is gaming sufficient for learning? Education for Information, **18** (4), 289-311.* ----------
- *Hansen, P. (1999) User interface design for IR interaction: a task-oriented approach. In: Aparac, T., Saracevic, T., Ingwersen, P. & Vakkari, P., eds. Proceedings of CoLIS 3, Third International Conference on the Conceptions of Library and Information Science: Digital Libraries: Interdisciplinary concepts, challenges and opportunities. Dubrovnik, 1999. Zagreb: Zavod za informacijske studije Odsjeka za informacijske znanosti: Filozofski fakultet; Lovke: Naklada Benja, pp. 191-205.*
- *Harter, S.P. (1992) Psychological relevance and information science. Journal of the American Society for Information Science, **43**(9), 602-615.*
- *Harter, S.P. (1996) Variations in relevance assessments and the measurement of retrieval effectiveness. Journal of the American Society for Information Science, **47** (1), 37-49.*

*Henninger, S. (1994) Using iterative refinement to find reusable software. IEEE Software, **11** (5), 48-59.*
- *Hersh, W., Pentecost, J. & Hickam, D. (1996) A task-oriented approach to information retrieval evaluation. Journal of the American Society for Information Science, **47** (1), 50-56.*
- *Hillman, D.J. (1964) The notion of relevance (1). American Documentation, **15** (1), 26-34.*
- *INEX (2002) http://qmir.dcs.qmw.ac.uk/INEX/index.html*
- *Ingwersen, P. (1992) Information retrieval interaction. London: Taylor Graham.*
- *Ingwersen, P. (1996) „Cognitive perspectives of information retrieval interaction: elements of a cognitive IR theory. Journal of Documentation, **52** (1), 3-50.*
- *Järvelin, K. & Kekäläinen, J. (2000) IR evaluation methods for retrieving highly relevant documents. In: Belkin, N.J., Ingwersen, P. & Leong, M-K., eds. Proceedings of the 23rd ACM Sigir Conference on Research and Development of Information Retrieval, Athens, Greece, 2000. New York, N.Y.: ACM Press, pp. 41-48.*
- *Jose, J.M., Furner, J. & Harper, D.J. (1998) Spatial querying for image retrieval. In: Croft, B.W., Moffat, A., van Rijsbergen, C.J., Wilkinson, R. & Zobel, J., eds. Proceedings of the 21st ACM Sigir Conference on Research and Development of Information Retrieval. Melbourne, 1998. ACM Press/York Press, pp. 232-240.*
- *Kekäläinen, J. & Järvelin, K. (2002) Evaluating information retrieval systems under the challenges of interaction and multidimensional dynamic relevance. In: Bruce, H., Fidel, R., Ingwersen, P. & Vakkari, P., eds. Emerging Frameworks and Methods, Seattle, 2002. Colerado: Libraries Unlimited, pp. 253-270.*
- *Kuhlthau, C.C. (1993) Seeking meaning: a process approach to library and information science. Norwood, NJ: Ablex Publishing.*
- *Lancaster, W.F. (1969) Medlars: report on the evaluation of its operating efficiency. American Documentation, **20** (2), 119-142.*
- *Lancaster, W.F. & Warner, A.J. (1993) Information retrieval today. Arlington: Information Resources Press.*
- *Losee, R.M. (1996) Evaluating retrieval performance given database and query characteristics: analytical determination of performance surfaces. Journal of the American Society for Information Science, **47** (1), 95-105.*
- *Losee, R.M. (1998) Text retrieval and filtering: analytical methods of performance. Norwell, Massachusetts: Kluwer Academic Publishers.*
- *Nielsen, M.L. (2002) The word association method: a gateway to work-task based retrieval. Åbo Akademi University Press, Åbo. Doctoral dissertation: Åbo Akademi University.*
- *Park, T.K. (1993) The nature of relevance in information retrieval: an empirical study. Library Quarterly, **63** (3), 318-351.*
- *Preece, J. et al. (1994) Human-computer interaction. Wokingham, England: Addison Wesley.*
- *Rasmussen, J., Pejtersen, A.M. & Goodstein, L.P. (1994) Cognitive systems engineering. N.Y.: John Wiley & Sons.*
- *Reid, J. (1999) A new, task-oriented paradigm for information retrieval: implications for evaluation of information retrieval systems. In: Aparac, T., Saracevic, T., Ingwersen, P. & Vakkari, P., eds. Proceedings of CoLIS 3, Third International Conference on the Conceptions of Library and Information Science: Digital Libraries: Interdisciplinary concepts, challenges and opportunities. Dubrovnik 1999. Zagreb: Zavod za informacijske studije Odsjeka za informacijske znanosti: Filozofski fakultet; Lovke: Naklada Benja, pp. 97-108.*
- *Reid, J. (2000) A task-oriented non-interactive evaluation methodology for information retrieval systems. Information Retrieval, **2** (1), 113-127.*
- *Robertson, S.E. (1981) The methodology of information retrieval experiment. In: Sparck Jones, K. ed., Information retrieval experiments. London: Butterworths, pp. 9-31.*
- *Robertson, S.E. & Hancock-Beaulieu, M.M. (1992) On the evaluation of IR systems. Information Processing & Management, **28** (4), 457-466.*
- *Robins, D. (1997) Shifts of focus in information retrieval interaction. In: Schwartz, C. & Rovrig, M., ed. Proceedings of the ASIS annual meeting (34). Silver Spring, Maryland, pp. 123-134.*
- *Rodden, K., Basalaj, W., Sinclair, D. & Wood, K. (2001) Does organisation by similarity assist image browsing? In: Proceedings of Human Factors in Computing Systems (CHI 2001), Seattle, WA. ACM Press, pp. 190-197.*
- *Rorvig, M. (1999) Images of similarity: a visual exploration of optimal similarity metrics and scaling properties of TREC topic-document sets. Journal of the American Society for Information Science, **50** (8), 639-651.*
- *Ruthven, I. (2001) Abduction, explanation and relevance feedback. University of Glasgow. Doctoral dissertation. Technical report: TR-2002-115.*
- *Ruthven, I., Lalmas, M. & van Rijsbergen, K. (2002) Ranking expansion terms with partial and ostensive*

evidence. In: Bruce, H., Fidel, R., Ingwersen, P. & Vakkari, P., eds. Emerging Frameworks and Methods, Seattle, 2002. Colerado: Libraries Unlimited, pp.199-219.

- *Saracevic, T. (1984) Measuring the degree of agreement between searchers. In: Flood, B, Witiak, J. & Hogan, T.H., eds. Proceedings of the 47th ASIS annual meeting. White Plains, NY, pp. 227-230.*
- *Saracevic, T. (1995) Evaluation of evaluation in information retrieval. In: Fox, E.A., Ingwersen, P., & Fidel, R., eds. Proceedings of the 18th ACM Sigir Conference on Research and Development of Information Retrieval. Seattle, 1995. N.Y.: ACM Press, pp. 138-146.*
- *Saracevic, T. (1996) Relevance reconsidered '96. In: Ingwersen, P. & Pors, N.O., eds. Proceedings of CoLIS 2, Second International Conference on Conceptions of Library and Information Science: Integration in Perspective. Copenhagen 1996. Copenhagen: Royal School of Librarianship, pp. 201-218.*
- *Schamber, L. (1994) Relevance and information behavior. In: Williams, M.E., ed. Annual Review of Information Science and Technology (ARIST) (29). Medford, NJ: Learned Information, INC. pp. 3-48.*
- *Schamber, L. Eisenberg, M.B. & Nilan, M.S. (1990) A re-examination of relevance: toward a dynamic, situational definition. Information Processing & Management, **26** (6), 755-775.*
- *Spink, A., Greisdorf, H., & Bateman, J. (1998) From highly relevant to not relevant: examining different regions of relevance. Information Processing & Management, **34** (5), 599-621.*
- *Stephen, P. & Hornby, S. (1997) Simple statistics: for library and information professionals. 2nd edition. London: Library Association Publishing.*
- *Su, L.T. (1992) Evaluation measure for interactive information retrieval. Information Processing & Management, **28** (4), 503-516.*
- *Swanson, D.R. (1977) Information retrieval as a trial-and-error process. Library Quarterly, **47** (2), 128-48.*
- *Swanson, D.R. (1986) Subjective versus objective relevance in bibliographic retrieval systems. Library Quarterly, **56** (4), 389-398.*
- *Tague-Sutcliffe, J. (1992) The pragmatics of information retrieval experimentation, revisited. Information Processing & Management, **28** (4), 467-490.*
- *Tang, R. & Solomon, P. (1998) Towards an understanding of the dynamics of relevance judgments: an analysis of one person's search behavior. Information Processing & Management, **34** (2/3), 237-256.*
- *TARGET on Dialog: 'how-to' guide. (1993). Palo Alto, USA: Dialog, 10 p.*
- *Taylor, R.S. (1968) Question negotiation and information seeking in libraries. College and Research Libraries, **29** (3), 178-194.*
- *Uhrskov, U.F. (2002) Er der forskel i søgeadfærd mellem humaniora- og naturvidenskabsstuderende? Biblioteksarbejde, **22** 63/1, 5-19.*
- *Urquhart, C. (1999) Using vignettes to diagnose information seeking strategies: opportunities and possible problems for information use studies of health professionals. In: Wilson, T.D. and Allen, D.K., eds. Proceedings of the 2nd international conference on research in information needs, seeking and use in different contexts. Sheffield, UK, 1998. London: Taylor Graham, pp. 277-289.*
- *Urquhart, C. (2001) Bridging information requirements and information needs assessment: do scenarios and vignettes provide a link? Information Research, **6** (2). http://www.informationr.net/ir/6-2/paper102.html*
- *Vakkari, P. (1999) Task complexity, problem structure and information actions: integrating studies on information seeking and retrieval. Information Processing & Management, **35** (6), 819-837.*
- *Vassileva, J. (1995) Ensuring a task-based individualized context for information retrieval from multimedia information systems for hospitals. In: Proceedings of IJCAI-95 Workshop on Intelligent Multimedia Information Retrieval, Montreal, 1995, pp. 172-185.*
- *Vassileva, J. (1996) A task-centered approach for user modelling in a hypermedia office documentation system. User Modelling and User Adapted Interaction, **6** (2-3), 185-223.*
- *Voorhees, E. (2001) Evaluation by highly relevant documents. In: Croft, W.B., Harper, D.J., Kraft, D.H. & Zobel, J., eds. Proceedings of the 24th ACM SIGIR Conference on Research a nd Development of Information Retrieval. New Orleans, LA, 2001. New York, N.Y.: ACM Press, pp. 41-48.*
- *Walker, S. (1989) The Okapi online catalogue research projects. In: Hildreth, C.R., ed. The Online catalogue: developments and Directions. London: The Library Association, pp. 84-106.*
- *Wersig, G. (1971) Information - kommunikation - dokumentation: ein beitrag zur orientierung der informations- dokumentationswissenschaften. München-Pullach: Verlag Dokumentation Saur KG.*
- *White, R., Jose, J. & Ruthven, R. (2001) Query-biased web page summarisation: a task-oriented evaluation. In: Croft, W.B., Harper, D.J., Kraft, D.H. & Zobel, J., eds. Proceedings of the 24th ACM SIGIR Conference on Research and Development of Information Retrieval. New Orleans, LA, 2001. New York, N.Y.: ACM Press, (poster), pp. 412-413.*
- *White, R., Ruthven, I. & Jose, J. (2002) Finding relevant documents using top ranking sentences: an*

evaluation of two alternative schemes. In: Beaulieu, M., Baeza-Yates, R., Myaeng, S.H. & Järvelin. K., eds. Proceedings of the 25th ACM SIGIR Conference on Research and Development of Information Retrieval. Tampere, Finland, 2002. New York, N.Y.: ACM Press, pp. 57-64.

- *Wilson, P. (1973) Situational relevance. Information Storage and Retrieval,* ***9*** *(8), 457-469.*
- *Wilson, T. D. (1999) Exploring models of information behaviour: the 'uncertainty' project. Information Processing & Management,* ***35*** *(6), 839-849.*
- *Winograd, T & Flores, C.F. (1986) Understanding computers and cognition. Norwwood, NJ: Addison-Wesley.*

How to cite this paper:

Borlund, Pia (2003) "The IIR evaluation model: a framework for evaluation of interactive information retrieval systems"Information Research, ***8****(3), paper no. 152 [Available at: http://informationr.net/ir/8-3/paper152.html]*

Contents **Home**