Invited Paper

JISTaP http://www.jistap.org
Journal of Information Science Theory and Practice

J. of infosci. theory and practice 1(3): 12-32, 2013
http://dx.doi.org/10.1633/JISTaP.2013.1.3.2

Interactive Information Retrieval: An Introduction

Pia Borlund*

Royal School of Library and Information Science
University of Copenhagen, Denmark
E-mail: sjc900@iva.ku.dk

ABSTRACT

The paper introduces the research area of interactive information retrieval (IIR) from a historical point of view. Further, the focus here is on evaluation, because much research in IR deals with IR evaluation methodology due to the core research interest in IR performance, system interaction and satisfaction with retrieved information. In order to position IIR evaluation, the Cranfield model and the series of tests that led to the Cranfield model are outlined. Three iconic user-oriented studies and projects that all have contributed to how IIR is perceived and understood today are presented: The MEDLARS test, the Book House fiction retrieval system, and the OKAPI project. On this basis the call for alternative IIR evaluation approaches motivated by the three revolutions (the cognitive, the relevance, and the interactive revolutions) put forward by Robertson & Hancock-Beaulieu (1992) is presented. As a response to this call the 'IIR evaluation model' by Borlund (e.g., 2003a) is introduced. The objective of the IIR evaluation model is to facilitate IIR evaluation as close as possible to actual information searching and IR processes, though still in a relatively controlled evaluation environment, in which the test instrument of a simulated work task situation plays a central part.

Keywords: interactive information retrieval, IIR, evaluation, human-computer information retrieval, HCIR, IIR evaluation model, user-oriented information retrieval, information retrieval, IR, history

1. INTRODUCTION

Interactive information retrieval (IIR), also known as human-computer information retrieval (HCIR) (Marchionini, 2006), concerns the study and evaluation of users' interaction with IR systems and their sat-

Open Access

Accepted date: September 12, 2013
Received date: September 6, 2013

***Corresponding Author:** Pia Borlund
Professor
Royal School of Library and Information Science
University of Copenhagen, Denmark
E-mail: sjc900@iva.ku.dk

isfaction with the retrieved information. 'Interactive' implies the involvement of human users in contrast to the notion of information retrieval (IR) only, which points to the system-oriented approach to IR signified with the Cranfield model (Cleverdon & Keen, 1966; Cleverdon, Mills & Keen, 1966). This approach is also referred to as TREC style evaluation (e.g., Belkin, 2008). Cool and Belkin (2011, p. 1) explain how the distinction between IR and IIR is found in the early history of computerized IR systems, especially in regard to their evaluation as carried out by different disciplinary groups. The history of IR evaluation can be traced back to 1953, and constitutes the origin of IR research as an empirical discipline with the series of tests that led to the Cranfield model (e.g., Ellis, 1996; Swanson, 1986). The different disciplinary groups mentioned by Cool and Belkin (2011, p. 1) are those of Computer Science and Library and Information Science addressing IR and IIR, respectively.

A user-oriented approach to IR has always existed in parallel to the system-oriented approach to IR evaluation, but the establishing of the IIR research area is relatively recent and takes its start in the early 1990s. Robertson and Hancock-Beaulieu (1992) explain the change and shift in focus that led to the establishing of the research area of IIR with the presentation of the three revolutions: the cognitive revolution, the relevance revolution, and the interactive revolution. At the same time Ingwersen (1992) published his book *Information Retrieval Interaction*, which points to the same tendency, namely a shift in focus from system-oriented IR towards IIR.

The objective of the present paper is to introduce the research area of IIR. We take an explicit focus on 'evaluation', because much research in IR deals with IR evaluation methodology due to the core research interest in IR performance, system interaction, and satisfaction with retrieved information (Järvelin, 2011, p. 113). Further, we approach the evaluation focus from a historical perspective. The idea is that we better understand IIR when we know where it comes from. For that reason we start out by introducing, in Section 2, the series of tests that led to the Cranfield evaluation model, which IIR is an alternative to. Hereafter Section 3 presents three iconic user-oriented studies and projects that all have had impact on how IIR is perceived and understood today: The MEDLARS test, the Book House fiction retrieval system, and the OKAPI project. The three revolutions by Robertson and Hancock-Beaulieu (1992) are presented in Section 4, which stipulates the modern and contemporary requirements to IIR evaluation. Section 5 presents current IIR evaluation by introducing the IIR evaluation model that uses simulated work task situations as an instrument of testing. Section 6 holds concluding statements and presents further readings to the research area of IIR.

2. THE CRANFIELD MODEL

As stated in the Introduction section the evaluation tradition of IR systems can be traced back to 1953, when two separate tests were carried out in the USA and UK, respectively. These were the Uniterm-tests of the Armed Services Technical Information Agency (ASTIA, USA) (Gull, 1956), and Cranfield (UK) (Cleverdon, 1960; Thorne, 1955). The research teams were evaluating the performance of the UNITERM system (developed by Mortimer Taube) against more conventional approaches to subject indexing and retrieval. These two tests became the first in a series of pioneering performance tests of indexing systems that led to the development of the Cranfield evaluation model. Central to this development is Cyril Cleverdon, who was the librarian of the College of Aeronautics at Cranfield (later the Cranfield Institute of Technology and Cranfield University). Cleverdon followed closely the USA-based ASTIA-Uniterm test, and he headed the UK tests. The UK test was designed building on the experiences of the ASTIA-Uniterm test. The Cranfield model and the related series of tests have over the years been described and discussed by numerous scholars (e.g., Ellis, 1996; Robertson, 1981; Sanderson, 2010; Sharp, 1964; Sparck Jones, 1981b; Swanson, 1965) in addition to the original publications (Cleverdon, 1960; Cleverdon, 1962; Cleverdon & Keen, 1966; Cleverdon, Mills & Keen, 1966; Gull, 1956; Thorne, 1955).

The two initial tests, the ASTIA-Uniterm team and the Cranfield-Uniterm team, differed from each other with respect to the employed evaluation methodology. In the case of the ASTIA-Uniterm test, two groups of people were involved. One group consisted of the indexing staff of ASTIA, and the second group consisted of staff from Mortimer Taube's company, Documentation Inc. The basic idea of the experiment was to use the *same* document collection–The existing ASTIA collection of 15,000 documents. Then each group indexed the document collection employing their own indexing system. The ASTIA group indexed the documents employing the operational ASTIA alphabetical subject heading list, and the staff from Documentation Inc. used the Uniterm system. Both groups then searched the document collections by the same 98 requests, which had been submitted to ASTIA in the normal course of its activities, and then they compared the relevance of the retrieved documents. The relevance definition employed by the groups was that of relevance to the request. The ASTIA team retrieved 2,220 documents that they considered relevant. The staff of Documentation Inc. retrieved 1,560 documents that they thought relevant. The retrieved document overlap in common to the two groups was only 580 documents. They were able to agree that 1,390 documents were relevant to the 98 requests. Of the total number of 3,780 documents retrieved by the two groups 1,577 documents were considered relevant by one group but irrelevant by the other. The two groups were unable to come to an agreement over which documents were relevant to which requests. Further, no decision procedure was made to help resolve their differences. The breakdown of the tests by the disagreements of the groups on the definition of relevance is to be considered the 'birth' of the relevance discussion to IR. Unfortunately, the relevance discussion within the groups was not given full-hearted attention, probably due to the groups' possible interest in the performance of each of their indexing systems. Ellis (1996, p. 2) comments on this by saying:

> It is unfortunate that the form of the dispute between the two groups is overshadowed by the fact that neither of the two parties could be said to be entirely disinterested in the outcome of the tests. Obviously, the ASTIA indexers would not look kindly on a system such as Uniterm which effectively transforms the, arguably, intellectual task of assigning subject headings to a document into the entirely clerical procedure of extracting keywords from a document's title or abstract. On the other hand, the staff of Documentation Incorporated were unlikely to favour any result which did not demonstrate the superiority of their product, particularly if the potential for prestigious and lucrative military contracts might depend on that result.

Later that year the Cranfield-Uniterm test was carried out in the UK (Cleverdon, 1960; Thorne, 1955). Basically, the Cranfield-Uniterm experiment had the same objective as that of the ASTIA-Uniterm test, that is, to compare the performance of the Uniterm system against a more conventional indexing system. However, the Cranfield team employed a different evaluation methodology in their experiment. Instead of using the actual document collection, a limited collection of 200 documents on the subject of aeronautics was set up, i.e. the dawn of the 'test collection' concept. From this collection a selection of 'source' documents was chosen. Based on the source documents 40 artificial requests were generated, to which the individual source documents represented the answers. The 40 requests were searched on the sub-collection of the 200 documents. Relevance was defined as the success in retrieving the source documents from which the corresponding requests were generated. In this way the Cranfield team avoided a relevance discussion like the one that ruined the ASTIA-Uniterm test. The main criticisms of the methodology was in regard to the employment of source documents. Ellis (1996, p. 3) summarises the criticisms as follows:

1. Only a single figure performance value was obtained (based on retrieval of the source documents); no corresponding figures were obtained for the retrieval of non-source documents, whether those might be considered 'relevant' or 'irrelevant'.
2. A-term based system might be said to be favoured over a concept-based system if the derivation of the artificial requests was influenced by the terms used in the source documents.

The Cranfield-Uniterm test functioned as a pilot experiment to the succeeding Cranfield I and Cranfield II tests. The Cranfield I test (Cleverdon, 1962) which was concerned with the comparative performance of four different indexing systems and made use of the principle of source documents, was carried out on a much larger scale than the initial Cranfield-Uniterm test (size of collection: 18,000 documents; number of requests: 1,200). In addition to the Cranfield-Uniterm test, the searches which failed to retrieve the source documents were subsequently analysed, in order to locate the failure of the retrieval: the request, the searching, the indexing, or the system.

The Cranfield I test was met with extensive criticism due to the dual role of the employment of source documents. Further, Swanson (1965, p. 6) points out that in an operational situation a source document generally does not exist, and that the relationship between the source document and the query was too close. Sharp (1964) recommends that "the source-document principle should be dropped and future tests carried out on the basis of taking into account *all* relevant documents retrieved" (Sharp, 1964, p. 174).[1]

The principle of source documents was abandoned in the Cranfield II test. The Cranfield II test (Cleverdon & Keen, 1966; Cleverdon, Mills & Keen, 1966), was a test of 33 different types of indexing languages, which varied in terminologies and structures. The test was based on a document collection consisting of 1,400 documents and 211 generated requests. Prior to the search, relevance of the documents to the search requests was determined. This was done in a two-step procedure. First, students of aeronautics searched the document collection and identified the relevant documents. Hereafter the relevant documents were sent to the generators of the given requests for proofing. The retrieval performance was defined as the retrieval of documents which had previously been identified as relevant to the request and measured according to the ratios of recall and precision.[2]

The described series of tests are worthy of consideration for two reasons: partly because the difficulties that were experienced in carrying them out influenced the methodology of subsequent tests, and partly because the features of these tests are embodied in today's main IR evaluation model, the Cranfield model. The Cranfield model derives directly from Cranfield II 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.

The Cranfield model constitutes the empirical research tradition on the development and testing of IR systems employed by the system-oriented approach to IR. The emphasis in this research tradition is on controlled laboratory tests. The objective of the employment of the Cranfield model is to keep all variables controlled and to obtain results which can state conclusions about retrieval systems in general. However, from an IIR point of view the Cranfield model suffers from limitations due to its restricted assumptions on the cognitive and behavioural features of the environment in which interaction with IR systems takes place. But with all respect it should be noted that Cleverdon and his colleagues were not blind to the user-oriented side of IR, which is seen in their comment on how the ideal evaluation of IR systems would be to use "actual questions with a relevance assessment made at the time by the questioner from complete texts" (Cleverdon, Mills & Keen, 1966, p. 15). However, at that time this was not possible. The MEDLARS test by Lancaster (1969), presented in Section 3.1, is an illustrative example of what was possible, but also of how information searching of scientific information took place in the mid-1960s.

3. USER-ORIENTED IR

This section introduces three iconic user-oriented IR studies and projects which reflect the different approaches to user-oriented IR that have formed IIR evaluation as of today: the MEDLARS test, the Book

[1] Sharp's recommendation is based on a review of the Cranfield-Western Reserve University test (Aitchison & Cleverdon, 1963) which was carried out at the same time as the Cranfield I test, employing the same evaluation methodology. This test is known also as Cranfield I½.

[2] Recall is the ratio of relevant documents retrieved over the total number of relevant document in the collection. Precision is the ratio of relevant documents retrieved over the total number of documents retrieved (Aitchison & Cleverdon, 1963, p. XII). Precision was originally called relevance, but was renamed from Cranfield II in order to avoid confusion with the general concept of relevance (Cleverdon, Mills & Keen, 1966, pp. 9-18).

House system, and the Okapi project. The MEDLARS test is interesting in that it is contemporary to that of Cranfield II, and as Cool and Belkin (2011, p. 10) put it the test should be well remembered for its attention to the human dimensions of failure analysis in IR systems. The MEDLARS test (Lancaster, 1969) can be viewed as top-down evaluation of an existing system in contrast to the Book House project, which took a bottom-up approach in the development of the Book House fiction retrieval system via a series of user studies, e.g., of user fiction requests (Pejtersen, 1980; Pejtersen & Austin, 1983; 1984), picture associations (Pejtersen, 1991), and search strategies for fiction literature (Pejtersen, 1992; Rasmussen, Pejtersen & Goodstein, 1994). With the introduction of online public access catalogues (OPACs) in the mid-1970s (in effect, an end-user IR system)—this type of research and system development came into focus. OPACS are in reality the first type of end-user IIR systems. The Book House project was about the development of a coherent end-user supporting OPAC. In addition, to be fully comprehensive in its empirical foundation of the system, the Book House system is a relevant innovation because it was ahead of its time and shows how icon-based fiction retrieval and the information needs of ordinary users were addressed. Similar to the Book House research the foci of the Okapi project were concerned with that of OPACs and their use. The research foci of Okapi can at a general level, in line with that of the Book house system, be categorised into three areas: IR functionality, navigation, and search strategy implementation; interface issues; and information searching behaviour and use. But more importantly Okapi is a showcase example of how system-oriented and user-oriented IR systems evaluations complement each other with the Okapi team's involvement in TREC from the very beginning. No doubt the work and experiences from the Okapi research have contributed to the recognition of the three revolutions presented in Section 4.

3.1 The MEDLARS Test

The present MEDLARS review is a description of the evaluation carried out and reported on by Lancaster (1969).[3] The MEDLARS system (now known as Medline[4] and PubMed[5]) has been evaluated by others beside Lancaster (1969), for instance by Salton (1972) in his comparison of the conventional indexing of MEDLARS and automatic indexing by the SMART system.[6] The reader should note that MEDLARS was not an end-user system at the time of testing; instead information search specialists searched on behalf of the user. At this time a question for information was put in writing and mailed to the National Library of Medicine and the questioner would wait for a reply (Robertson & Hancock-Beaulieu, 1992, p. 460). Lancaster (1969, p. 119) describes MEDLARS as "a multipurpose system, a prime purpose being the production of *Index Medicus* and other recurring bibliographies." The aim of Lancaster's MEDLARS test was to evaluate the existing system and to find out ways in which the performance of the system could be improved.

The planning of the MEDLARS evaluation began in December 1965 and was carried out during 1966 and 1967. Lancaster (1969, p. 120) outlines the principal objectives of the evaluation as: 1) to study the demanded types of search requirements of MEDLARS users; 2) to determine how effectively and efficiently the present MEDLARS service meets these requirements; 3) to recognise factors adversely affecting performance; and 4) to disclose ways in which the requirements of MEDLARS users may be satisfied more efficiently and/or economically.

Lancaster (1969, p. 120) further explains how the users were to relate the prime search requirements to the following factors:

1. The *coverage* of MEDLARS (i.e., the proportion of the useful literature on a particular topic that is indexed into the system, within the time limits imposed).

[3] MEDLARS is an acronym for MEDical Literature Analysis and Retrieval System.

[4] Medline URL: http://medline.cos.com/

[5] PubMed URL: http://www.ncbi.nlm.nih.gov/pubmed/

[6] SMART is an acronym for System for the Mechanical Analysis and Retrieval of Text.

2. Its *recall* power (i.e., its ability to retrieve "relevant" documents, which within the context of this evaluation means documents of value in relation to an information need that prompted a request to MEDLARS).
3. Its *precision* power (i.e., its ability to hold back "non-relevant" documents).
4. The *response time* of the system (i.e., the time elapsing between receipt of a request at a MEDLARS centre and delivery to the user of a printed bibliography).
5. The *format* in which search results are presented.
6. The amount of *effort* the user must personally expend in order to achieve a satisfactory response from the system.

The document collection used was the one currently available on the MEDLARS service at the time, which according to Robertson (1981, p. 20) had the size of approximately 700,000 items. 302 real information requests were used to search the database, and the users requesting the information made the relevance assessments. A recall base was generated based on a number of documents judged relevant by the users in response to their own requests, but found by means outside MEDLARS. The sources for these documents were (a) those already known to the user, and (b) documents found by MEDLARS staff through sources other than MEDLARS or *Index Medicus*. Subsequently, each user was asked to assess relevance of a sample of the output from the MEDLARS search, together with selected documents from other sources (Robertson, 1981, p. 20). The relevance assessments were carried out with reference to three categories (degrees) of relevance: major, minor, or no value. In addition, the users explained their relevance judgements by indicating why particular items were of major, minor, or no value to their information need (e.g., see Martyn & Lancaster, 1981, p. 165 for an example of the relevance questionnaire used in the MEDLARS test). Precision and relative recall ratios were calculated and compared to the degree of exhaustivity[7] and specificity[8] of the requests as expressed by use of the controlled entry vocabulary. Analysis of failures was carried out, and a classification of reasons for failure was devised (Lancaster, 1969, p. 125, Table 5).

Variables were built into the test. The primary variable concerned the level of interaction between the user and the system. The interaction could take place at three different levels, which refers to, i.e., the originality or level of influence on the expression of the user's information request (Lancaster, 1969, p. 120). At the first level the interaction takes place as a *personal interaction* that is, the user makes a visit to a MEDLARS centre and negotiates his requirements directly with an information search specialist. The second level of interaction is identified as *local interaction*, which means that the user's information request comes by mail to the centre, but is submitted by a librarian or information specialist on behalf of the requester. The third interaction level concerns *no local interaction*, which implies that no interference by a librarian or information specialist has taken place in the process of interpreting or translating the request into the controlled vocabulary used by MEDLARS. In other words, the request comes directly by mail from the requester. The different levels of interaction enabled a comparison of the results obtained. The main product of the test was a detailed analysis of the reasons for failure. The result is interesting, because "it appears that the best request statements (i.e., those that most closely reflect the actual area of information need) are those written down by the requester in his own natural-language narrative terms" (Lancaster, 1969, p. 138). For the user-oriented approach to IR systems evaluation this result is seen as a strong piece of evidence and motivation for exactly this type of evaluation. That only the user who owns the information need can represent and assess that information need hereby gives insight into how well the system works under real-life operational conditions in a given situation and context at the time the information is requested.

The results, the focus, and the approach of the MED-

[7] Exhaustivity and specificity are well-established concepts within indexing. Exhaustivity refers to the comprehensiveness of the indexing in question, which consequently affects retrieval results.

[8] Specificity refers to the accuracy of the indexing. Specificity of indexing makes for high precision and low recall.

LARS test helped contemporary system-driven researchers to understand problems of IR as well as to comprehend how to design future systems and tests. In the following quotation, Sparck Jones (1981a, p. 230) comments on the effect of Lancaster's test by describing the situation prior to the MEDLARS test:

> It was thus not at all obvious how systems should be designed to perform well, modulo a preference for recall or precision, in particular environments, especially outside established frameworks like those represented by the Medlars system, or for situations and needs clearly resembling those of existing systems.

Robertson (1981, pp. 20-21) describes how Cranfield II and MEDLARS are two of the classical tests, with Cranfield II being a highly controlled and artificial experiment, and MEDLARS being an investigation of an operational system, as far as possible under realistic conditions. They both play a significant part in creating the archetypes of tests, by each representing the opposite poles of the system-and user-oriented spectrum.

3.2. The Book House Fiction Retrieval System

The Book House system is a pictogram and icon based system designed for the retrieval of fiction literature. The development of the Book House system started in the 1970s and was headed by Annelise Mark Pejtersen. The initial development of the Book House system, including the work on the Book Automat, started while Pejtersen was employed at the Royal School of Librarianship, Denmark (now the Royal School of Library and Information Science). Later the research and development project of the Book House system followed her to Risø National Laboratory (near Roskilde, Denmark). Pejtersen (1992) describes the Book House system as "an interactive, multimedia, online public access catalogue [OPAC] designed to support casual novice users in information retrieval. It uses icons, text and animation in the display interface in order to enhance the utility of the system" (Pejtersen, 1992, p. 359). Further, the main characteristic of the system is its metaphoric use of the familiar (Danish) public library of which the user is invited into a virtual space version. The Book House metaphor is chosen in support of the user's memory and navigation by locating information about the functionality and content of the system in a coherent and familiar spatial representation (Rasmussen, Pejtersen & Goodstein, 1994, p. 290). In other words, the Book House system presents a spatial metaphor built around a library that houses fiction (Pejtersen, 1989, p. 40), and is not to resemble a bookshop, as indicated by Wilson (2011, pp. 145-146).

The pictorial interface of the Book House system deserves a brief description for the sake of the reader's visualisation and imagination of it. The interface of the Book House system, which the user meets first, is a picture of a house built of books with people entering it via the open front doors. The user accesses the Book House system by clicking with the computer mouse at the open front doors, and enters the hall of the house. The Book House consists of several rooms. Basically, the user 'walks' through the rooms with different arrangements of books and people (Rasmussen, Pejtersen & Goodstein, 1994, p. 289). The first room of the Book House is the hall. Here the user can choose between different library book collections represented by rooms (in fact different databases). These are the children's collection; the adults' collection; or a merged version of the previous two collections known as the 'family collection.' The type of collection is illustrated with the people in front of the entrances of the collection rooms, and put in writing on top of the entrances. In front of the children's collection children are pictured; similarly adults are positioned in front of the adults' collection; and together a child and a grown-up person are ready to enter the family collection. No matter what collection the user chooses the collection rooms look the same, with the only differences being the people inhabiting them and the size of the furniture. If the children's collection has been selected the user meets only children in that room and so on. Having chosen any of the three collection rooms, the next room contains a choice of search strategies. The choice of search strategy takes the user one step deeper into the Book House. To maintain an overview of where the user is, a horizontal breadcrumb trail is displayed at the top of the screen that shows all the steps taken, and the user can return to any previous step of the trail by clicking that icon.

The development of the Book House system and the evaluation of the concepts underlying the design, as well as the means chosen for implementation, have involved a considerable amount of empirical studies. Initially, a classification system for fiction literature was developed based on an empirical field study of users' information requests. The analysis of the requests revealed that basically fiction requests can be categorised into five so-called 'dimensions.' These are: 1) the subject-matter of the book (e.g., a mystery novel); 2) the frame of the book with reference to time and place (e.g., a historical novel); 3) the author's intention or attitude (e.g., a humorous book); 4) the accessibility of the book (e.g., an easy read book, or a book with big letters); and 5) 'other request formulations' (e.g., something like Emily Brontë) (Pejtersen, 1980, pp. 148-149). This part of the research is reported in, e.g., Pejtersen (1980) and Pejtersen and Austin (1983; 1984).

The research continued with the construction of a thesaurus based on users' word associations relying on the fiction classification system (Pejtersen, Olsen & Zunde, 1987). This further led to research into a picture association thesaurus for implementation into the Book House system (Pejtersen, 1991). Two tests were carried out before an actual implementation of the picture association thesaurus. Pejtersen (1991, p. 119) describes how the focus on the first test was primarily "to determine whether it was feasible for designers to associate pictures from keywords chosen from different dimensions of the classification scheme used in the database of the Book House and draw small pictures to represents these words". The second test was concerned with "whether a sufficient consensus could be achieved between designers' conception of associative relationships between pictures and keywords and that of different groups of potential users of the system" (Pejtersen, 1991, p. 119). With the picture association thesaurus implemented into the Book House the usability and effectiveness were tested in a real work context.

Pejtersen (1992, p. 362) further explains how it was possible, based on an interaction study of user-librarian negotiations, to verify five different search strategies. Of these five strategies four were used to design the navigation and the retrieval functionality of the Book House system. Briefly, the strategies are: 1) *bibliographic strategy*, when searching for a known item; 2) *analytical strategy*, the search for book attributes matching users' need; 3) *analogy strategy*, an associative type of search based on analogies between known books and unread books building on best-match algorithms; 4) *browsing strategy*, intuitive searching (browsing) of book shelves looking at book contents and book cover pictures for recognition of something of interest; and 5) *empirical strategy*, which is based on feedback and check routines, that is, the users' acceptance and rejection of documents proposed by the system based on a comparison match of books and the user's indication of his or her information need (Pejtersen, 1992, p. 362; Rasmussen, Pejtersen & Goodstein, 1994, pp. 275-76). It is the fifth strategy, the *empirical strategy*, which is not implemented into the Book House. The *empirical strategy* is commonly known and implemented in other systems as relevance feedback mechanisms. As such the research and development of the Book House, based on users' IR and searching behaviour, empirically verifies relevance feedback as an end-user search strategy.

The four implemented search strategies are visualised and illustrated in the room of the Book House where the user chooses the search strategy. The strategies are depicted as metaphors with a functional analogy to users' activities in a library. Four persons are viewed using the library and the strategies of browsing (that is, browsing pictures or browsing book descriptions), analogy search, and the analytical type of search. The bibliographic strategy is available as a sub-search strategy of the analytical strategy.

The Book House system itself was evaluated in an operational setting of the public library of Hjortespring in Denmark over a six-month period in 1988. The test participants were the actual users of the public library, children as well as adults, ranging from the age of seven to 70, and resulting in 2850 search logs and 75 complete user observations (e.g., Goodstein & Pejtersen, 1989, p. 168).

As a development and research project the Book House system is unique. Every single test involved operational conditions and test participants, and even the so-called laboratory tests of the implemented search strategies made use of test participants (librarians were invited to the laboratory). Briefly summarised, the Book House research project counts the following essential contributions: the development of a classification system for fiction literature; the development and construction of a picture association thesaurus; the verification of library end-user search strategies as well as the implementation of these strategies; and the development of a transparent, analogy, and icon based interface (at a time when the use of icons was not an interface tradition). The interface is self-explanatory through the visualisation of the public library environment, including the presentation of library activities, which the users can recognise without problems. The Book House system is also to be seen as a revolutionary new type of OPAC. The problem with the traditional OPACs was that they merely functioned as an electronic version of the traditional card catalogue; secondly, most OPACs reflected primarily the librarians' use and need for a work tool, not the users' need for a tool to direct them to the information objects. In these respects the Book House system made a difference as it was based on observations and field studies of end-users' use of libraries. It provided the users with an OPAC, made for the users, presented in the frame and context of the familiar public library environment. Another important contribution by the Book House development is the gained experience on *cognitive domain analyses*. These experiences have been assembled together in a very extensive and comprehensive framework for work (task) centred evaluation and design (e.g., Pejtersen & Fidel, 1998; Pejtersen & Rasmussen, 1998; Rasmussen, Pejtersen & Goodstein, 1994). The framework model is due to its illustrative representation of the evaluation boundary levels, popularly referred to as the 'onion-model.' The evaluation framework is recently discussed in detail in the very recommendable book by Fidel (2012).

3.3. The OKAPI Project

Okapi is the name of a series of generations of test retrieval systems.[9] Okapi can be viewed as a test bed including an experimental test system (or versions of systems), and can as such be considered the user-oriented IR approach's answer to Salton's SMART project within the system-oriented IR evaluation approach (e.g., Salton, 1981). The foci of the system tests and the different installations of the systems are reported in a number of publications (e.g., Beaulieu & Jones, 1998; Beaulieu, Robertson & Rasmussen, 1996; Robertson & Hancock-Beaulieu, 1992; Robertson, 1997a; Walker & De Vere, 1990; Walker, 1989). Even an entire issue of the *Journal of Documentation* (Robertson, 1997b) has had the theme of Okapi and IR research. The Okapi systems have had homes at various locations. From 1982-1989 Okapi was based at the Polytechnic of Central London (now the University of Westminster). From 1989 Okapi has been located at City University, London.

Okapi is a test search system designed for end-users who are not expert searchers. Okapi has throughout its history been used as the basis for real services to groups of users. The initial Okapi investigations started out with Okapi functioning as an OPAC that was made available in a number of British libraries to the actual users of those libraries. Since then, various Okapi systems have been made available in similar ways to groups of researchers over a network with a database of scientific abstracts (Robertson, 1997a, p. 5). The operational setup of the Okapi was made specifically with the purpose to create an environment in which ideas could be subject to user trials. The operational setup of the Okapi investigations allowed for end-users to be observed in a large variety of ways, from straight transaction logs to questionnaires, interviews, and direct observation (Robertson, 1997a, p. 5). This means that the Okapi team received additional qualitative-based information on the functionality and performance of, e.g., relevance feedback. Robertson (1997a, p. 5) explains how the quality of user-oriented Okapi investigations supply evidence concerning relevance feedback, not just on its retrieval effectiveness in the usual sense, but also on the ways in which users actually make use of it, and how useful they find it. The Okapi project has contributed to the understanding that users can be given access to sophisticated IR and feedback techniques and that users are capable of using these techniques effectively. At the same time the research shows that there are difficulties involved in

[9] Okapi is an acronym for Online Keyword Access to Public Information.

providing these techniques, and that users have to be given guidance in the use of these techniques, as pointed out by Robertson (1997a, p. 6).

Robertson (1997a, pp. 3-4) elegantly describes the typical Okapi system and its underlying probabilistic-based retrieval and relevance feedback mechanisms:

> What the user sees first is an invitation to enter a query in free-text form. This free-text query is then parsed into (generally) a list of single word-stems; each stem is given a weight based on its collection frequency. The system then produces a ranked list of documents according to a best-match function based on the term weights, and shows the user titles of the top few items in the list. The user can scroll the list and select any title for viewing of the full record. Having seen the full record, he or she is asked to make relevance judgement ("Is this the kind of thing you want?") in a yes/no term. Once the user has marked a few items as relevant, he or she has the opportunity to perform relevance feedback search ("More like the ones you have chosen?"). For this purpose, the system extracts terms from the chosen documents and makes up a new query from these terms. This is normally referred to as query expansion, although the new query may not necessarily contain all the original terms entered by the user. The new query is run and produces a ranked list in the usual fashion, and the process can iterate.

The quotation by Robertson regarding a typical Okapi system session reveals the multi-layer nature of the system architecture and use of the system, which is in accordance with the broad system definition in the present user-oriented IR evaluation approach. The broad definition is further reflected in the investigative and experimental division of the focus of Okapi tests on: IR system techniques, information searching behaviour, and interface issues. The inclusion of research on interface issues emphasizes the point of difference between the two main approaches to IR systems evaluation made by Ingwersen (1992, p. 87) about the question as to where IR systems end and the automatic interface intermediary begins. The interface research of Okapi concerns interaction issues such as the user's perception of system functions in relation to information searching tasks (e.g., Beaulieu, 1997; Beaulieu & Jones, 1998). Via Okapi, research attention is given to the issue of the 'cognitive load' that IR systems put on the users. Cognitive load refers to the intellectual effort required from the user in order to work the system. As such the research areas of IR and human computer interaction (HCI), especially with the introduction of OPACs and IIR systems, become overlapping. Beaulieu and Jones (1998, pp. 246-247) explain that cognitive load "involves not only the requirement to understand conceptual elements of the system [...] but to make meaningful decisions based on that understanding. When searchers have to make decisions without adequate understanding, their efforts can be counter-productive for overall system effectiveness." This they illustrate with an example of how users who do not realize how query expansion operates may avoid the trouble of making relevance judgements if they can, and lose one important advantage of probabilistic retrieval. So the ideal interface for an IIR system is one that meets the (inexperienced) end-user as transparent and self-explanatory so that no extra cognitive burden is put on the user–like the Book House system.

The Okapi project further serves as a very good example of how the two main approaches to IR systems evaluation complement each other. In that, Okapi has taken part in several TREC[10] tracks over the years, e.g., ad hoc, routing, interactive, filtering, and web tracks. TREC is an annual research workshop and provides for large-scale test collections and methodologies, and builds on the Cranfield model (Voorhees & Harman, 2005a). TREC is hosted by the U.S. Government's National Institute of Standards and Technology (NIST). According to Robertson, Walker, and Beaulieu (1997, p. 20) the Okapi team uses TREC to improve some of the automatic techniques used in Okapi, specially the term weighting function and the

[10] TREC is an acronym for Text REtrieval Conference. For more information about TREC, the reader is directed to the book *TREC: Experiments and Evaluation in Information Retrieval* by Voorhees and Harman (2005b), as well as to the following URL: http://trec.nist.gov/.

algorithms for term selection for query expansion. Prior to TREC, the Okapi team had worked only with operational systems or with small-scale partially controlled experiments with real collections (Robertson, Walker & Beaulieu, 1997, p. 23). As such TREC is a change of test culture and environment to the Okapi test tradition. The Okapi team maintains an interest in real users and information needs because to them the most critical and most difficult areas of IR system design are in the area of interaction and user interfaces (Robertson, Walker & Beaulieu, 1997, p. 32). The Okapi team sees their participation in the TREC experiments as complementary to the real user studies (Robertson, 1997a, p. 6). Further, Okapi's participation in TREC illustrates how systems, or elements of systems, at some point may be tested according to system-oriented principles, as well as how the interactive functionality of parts or more 'finished' and complete IIR systems are best evaluated and validated by involvement of end-users and potentially dynamic information needs. Okapi's participation in TREC also shows that the boundaries between the two evaluation approaches are not clear-cut.

The Okapi project represents a strong and illustrative case of user-oriented IR systems evaluations (e.g., Beaulieu & Jones, 1998; Beaulieu, Robertson & Rasmussen, 1996; Robertson & Hancock-Beaulieu, 1992; Robertson, 1997a; Walker & De Vere, 1990; Walker, 1989). Further, Okapi demonstrates in relation to TREC the complementary nature of the two main approaches to IR systems evaluation. In addition, Okapi and the series of tests in which probability-based IR techniques such as relevance feedback have been tested emphasise the need for alternative approaches to evaluation of IIR techniques and search facilities, which we focus on in the following section with the presentation of the three revolutions. Though the present introduction of Okapi is made from a historical viewpoint it is appropriate to point out that Okapi is still active and maintained. Okapi has been extended to handle XML documents and element retrieval for INEX[11] (Lu, Robertson & MacFarlane, 2006; Lu, Robertson & MacFarlane, 2007; Robertson, Lu & MacFarlane, 2006). The Okapi-Pack, which is a complete implementation of the Okapi system, is available from the Centre For Interactive Systems Research (CISR) at City University for a nominal fee when used for research purposes only (URL: http://www.soi.city.ac.uk/~andym/OKAPI-PACK/).

4. THE THREE REVOLUTIONS

In 1992 Robertson and Hancock-Beaulieu (1992) wrote a paper on the current status of the evaluation of IR systems. The paper was a follow-up of Robertson's chapter in the book *Information Retrieval Experiment*, edited by Sparck Jones (1981c) and dedicated to Cyril Cleverdon. For decades this book remained the one substantial book on the evaluation of IR systems. But with their follow-up paper Robertson and Hancock-Beaulieu present a call for alternative approaches to IR systems evaluation: that is, alternative with respect to the Cranfield model. They explain and illustrate the need for alternative approaches to IR systems evaluation with what they refer to as: the *cognitive revolution*; the *relevance revolution*; and the *interactive revolution*.

The *cognitive revolution* concerns the nature of an information need and its formation process. As a result of the *cognitive revolution* the information need is viewed as a reflection of an anomalous state of knowledge (ASK) (Belkin, 1980; Belkin, Oddy & Brooks, 1982) on the part of the requester. This refers to how an information need is understood and acknowledged as a dynamic and individual concept. This means that an information need, from the user's perspective, is a personal and individual perception of a given information requirement (Belkin et al., 1993), and that an information need for the same user can change over time (e.g., Ellis, 1989; Kuhlthau, 1993; Spink, Greisdorf & Bateman, 1998).

The *relevance revolution* points to the increasing acceptance that a stated request put to an IIR system is not the same as an information need, and therefore

[11] INEX is an acronym for INitiative for the Evaluation of XML retrieval. For more information about INEX the reader is directed to: URL: http://inex.is.informatik.uni-duisburg.de/

relevance should be judged in relation to the need rather than the request. In addition, there is a growing recognition of the multidimensional *and* dynamic nature of the concept of relevance (e.g., Borlund, 2003b; Schamber, Eisenberg & Nilan, 1990). The multidimensional nature of relevance, the fact that relevance is not an absolute quality, is empirically documented, for instance, by how relevance can be divided into classes and types of relevance, and by how the concept is applied with reference to various criteria and degrees and at different levels. A compiling list by Schamber (1994) of 80 relevance criteria that influence users' relevance judgements imply that the concept of relevance consists of many facets and therefore should not be treated as a binary variable (relevant/not relevant). The dynamic nature of relevance is demonstrated by, e.g., Bruce (1994), and Spink and colleagues (Spink, Greisdorf & Bateman, 1998) who found that users' relevance criteria may change over the course of session time.

The third and final revolution, the *interactive revolution*, points to the fact that IR systems have become interactive and consequently cannot be evaluated without including the interactive seeking and retrieval processes. IIR systems are defined as systems where the user dynamically conducts searching tasks and correspondingly reacts to system responses over session time. Thus, the foci of IIR system evaluation may include all the user's activities of interaction with the retrieval and feedback mechanisms as well as the retrieval outcome itself.

These 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 user request and search statement. This implies the assumption that learning and modifications by users are confined to the search statement alone. Furthermore, this model has a strong tradition for using only binary, topical relevance, ignoring the fact that relevance is a multidimensional and dynamic concept (Borlund, 2003b). 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. In essence this is about, on the one hand, control over experimental variables, observability, and repeatability, and on the other hand, realism (Robertson & Hancock-Beaulieu, 1992, p. 460). The three revolutions summarise the fundamental causes that have led to the current demand for alternative approaches to the evaluation of IIR systems.

5. IIR EVALUATION AS OF TODAY

The call for IIR evaluation approaches as presented by Robertson and Hancock-Beaulieu (1992) is supported by numerous scholars (e.g., Belkin, 2008; Ellis, 1996; Harter, 1996; Järvelin, 2011; Saracevic, 1995). Belkin (2008, p. 52) addresses in his 2008 ECIR Keynote the general need for more user-oriented IR research and in particular the need for alternative evaluation approaches to the Cranfield model. Belkin explicitly highlights *the IIR evaluation model* by Borlund (e.g., Borlund, 2003a), which employs simulated work task situations as the central instrument for testing, as such an attempt.

5.1 The IIR Evaluation Model

The IIR evaluation model meets the requirements of the three revolutions put forward by Robertson and Hancock-Beaulieu (1992) in that the model builds on three basic components: (1) the involvement of potential users as test participants; (2) the application of dynamic and individual information needs (real, and simulated information needs); and (3) the employment of multidimensional and dynamic relevance judgements. The cognitive revolution concerning the individual and dynamic nature of information needs is taken into account by allowing test participants to work with personal and individual information need interpretations of both their own and simulated information needs. The test participants' need interpretations are allowed to develop and mature over session time for the same test participant, a dynamic nature which is proved to be strongly connected to the process of assessing relevance (Borlund & Ingwersen, 1997). The relevance revolution is taken into account by having relevance judged in relation to the need– and in addition, in relation to the underlying situation of the need. Furthermore, the concept of relevance is in a way appropriate to its dynamic and multidimensional nature by being assessed interactively in a non-binary way (e.g., Borlund 2003b). The interactive revolution is incorporated into the IIR evaluation model by having the test participants work with personal information need interpretations which they try to satisfy through information searching and retrieval processes.

The aim of the IIR evaluation model is to facilitate IIR evaluation as close as possible to actual information searching and IR processes, though still in a relatively controlled evaluation environment. The IIR evaluation model builds on the concept of a stable simulated work task situation in order to frame the simulated information need situation, but allowing for its modification and development (by learning processes) during searching and retrieval. Both the information need and the derived request (or 'topic') are thus allowed to shift focus during the process. At the same time, the simulated work task situation acts as the point of reference against which situational relevance is measured. Situational relevance is understood as an assessment which points to the relationship between an information object presented to the user and the cognitive situation underlying the user's information need. As a generic concept, it may refer to the usefulness, usability, or utility of such objects in relation to the fulfilment of goals, interests, work tasks, or problematic situations intrinsic to the user. This understanding is in line with the interpretation proposed by Schamber, Eisenberg, and Nilan (1990). 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; *and*

Part 3. Alternative performance measures capable of managing non-binary based relevance assessments.[12]

Part 1 and 2 concern the collection of data, whereas Part 3 concerns data analysis. As such the IIR evaluation model is comparable to the Cranfield model's two main parts: the principle of test collections and the employment of recall and precision. The first part of the model deals with the experimental setting. This part of the model is identical to the traditional user-oriented approach in that it involves potential users as test participants, applies the test participants' individual and potentially dynamic information need interpretations, and supports assignment of multidimensional relevance assessments (that is, relevance assessment according to various types of relevance, several degrees of relevance, and multiple relevance criteria) and dynamic relevance assessments (allowing that the users' perception of relevance can change over time) (Borlund, 2003b). The IIR evaluation model differs from the traditional user-oriented approach with the introduction of simulated work task situations as a tool for the creation of simulated, but realistic, information need interpretations. Hence, Part 2 outlines recommendations for how to create and use simulated work task situations, presented in detail in Section 5.1.1. No doubt the major challenge is the design of realistic and applicable simulated work task situations.

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[13] that enables the facilitation of 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 are capable of managing non-binary based relevance assessments, as a result of the application of the Parts 1 and 2 of the model. The dominating

[12] With respect to the performance measures of recall and precision traditionally employed.

[13] 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.

use of the ratios of recall and precision for the measurement of the effectiveness of IR performance, also within the traditional user-oriented approach to IR systems evaluation, has forced researchers to reconsider whether these measures are sufficient in relation to the effectiveness evaluation of IIR systems. Spink and colleagues comment on the situation in the following way: "[t]he 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, Greisdorf & Bateman, 1998, p. 604). Nevertheless, the reason for the measures' well-established positions is due to the clear and intuitively understandable definitions combined with the fact that they represent important aspects of IR, are easy to use, and the results are comparable (Sparck Jones, 1971, p. 97). However, the measures view relevance as a binary concept and do not allow for the often non-binary approach taken in, e.g., the user-centred approaches. Further, the measures do not distinguish between the different types of relevance involved, but treat them as one and the same type. In order to illustrate the need for alternative performance measures, the measures of Relative Relevance (RR) and Ranked Half-Life (RHL) (Borlund & Ingwersen, 1998) were introduced, followed up by the stronger measures of cumulative gain (CG) with, and without, discount by Järvelin and Kekäläinen (2000). The RR measure is intended to satisfy the need for correlating the various types of relevance applied in the evaluation of IR and specifically IIR systems. The RHL informs about the position of the assessed information objects in regard to how well the system is capable of satisfying a user's need for information at a given level of perceived relevance. In line with RHL, the CG measures are positional measures. The assumptions of the CG measures are that:

- Highly relevant documents are more valuable than marginal ones; *and*
- The lower the ranked position of a retrieved document, the less valuable it is for the user, because the less likely it is that the user will ever examine it.

(Järvelin & Kekäläinen, 2000, p. 42)

The recommendation is to apply these measures in combination with the traditional performance measures of recall and precision. A discussion of the strengths and weaknesses of the RR, RHL, and CG measures can be found in the book by Ingwersen and Järvelin (2005).

5.1.1 The Test Instrument of a Simulated Work Task Situation.

The simulated work task situation is a short textual description that presents a realistic information-requiring situation that motivates the test participant to search the IR system (Borlund, 2003a). A simulated work task situation serves two main functions: 1) it causes a 'simulated information need' by allowing for user interpretations of the simulated work task situation, leading to cognitively individual information need interpretations as in real life; and 2) it is the platform against which situational relevance is judged by the test participant (Borlund & Ingwersen, 1997, pp. 227-228). More specifically it helps to describe to the test participants:

- 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 participants understand the objective of the search.

(Borlund & Ingwersen, 1997, pp. 227-228)

As such the simulated work task situation is a stable concept, i.e., the given purpose and goal of the IR system interaction. Figure 1 depicts a classic example of a simulated work task situation tailored towards university students.

Further, by being the same for all the test participants experimental control is provided, and the search interactions are comparable across the group of test participants for the same simulated work task situation. As such the use of simulated work task situations ensures the IIR study will possess both realism and control.

The issue of realism of the descriptions of the simulated work task situations is very essential in order for the prompted search behaviour and relevance assessments of the test participants to be as genuine as intended. Therefore realism is emphasised in the requirements of how to employ simulated work task situations (Borlund, 2003a). In brief, the requirements are as follows:

1) To tailor the simulated work task situations towards the information environment and the group of test participants;
2) To employ either a combination of simulated work task situations and indicative requests (simulated situations), or simulated work task situations only;
3) To employ both simulated work task situations and real information needs within the same test;
4) To permute the order of search jobs;
5) To pilot test; *and*
6) To display employed simulated work task situation when reporting the IIR study.

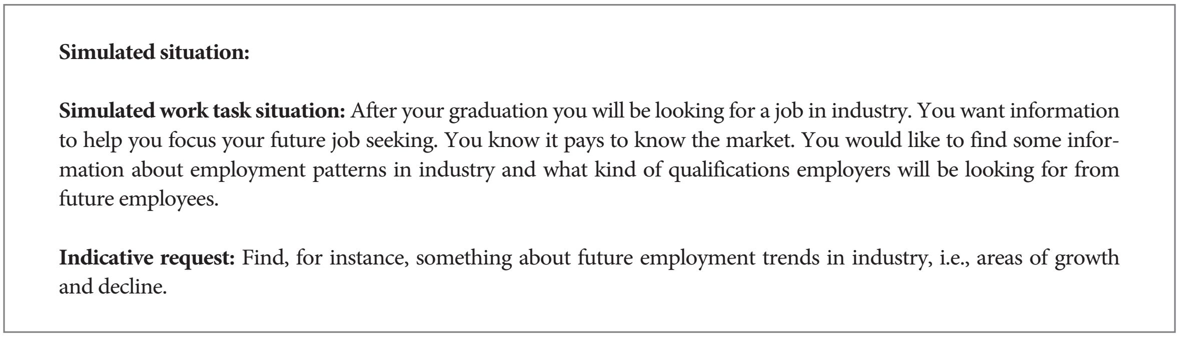
A well-designed simulated work task situation should be tailored to fit the type of searching under study and the group of test participants, and is one which:

- The test participants can relate to;
- They can identify themselves with;
- They find topically interesting; *and*
- The simulated work task situation must also provide enough imaginative contexts in order for the test participants to be able to relate to and apply the situation.

If the evaluation takes place by involvement of university students then the simulated work task situation should be to describe a situation they can relate to, which they can identify themselves as being in, and to present a topic of searching they find interesting. The described situation should be authentic, relevant, and realistic to the university students – males and females – so that it leads to realistic interpretations and interactions with the simulated information needs. The requirement to tailor the simulated work task situations entails homogeneity of the group of test participants. They need to have something in common, which can form the foundation for the design, tailoring, and use of the simulated work task situations.

The second requirement concerns evaluation by use of either a combination of simulated work task situations and indicative requests, or only simulated work task situations. This requirement provides for an option to direct the searching or help the test participants with determining what to search for. When testing one can decide to include or exclude the indicative request (see Figure 1). When included, the simulated work task situation is followed up with a suggestion of what to search for in the form of an indicative request. When excluded the simulated work task situation stands alone. Previous research (Borlund 2000a; 2000b) shows that test participants make use of the indicative request in different ways: to some the indicative request made it easier to generate the search formulations as they picked the search terms from the indicative requests. Another test participant revealed that the indicative request helped him understand what was expected from him. Yet another test participant explained that 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.

The third requirement concerns how to employ a combination of simulated work task situations (simulated information needs) and the test participants' gen-



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.

Fig. 1 Example of a simulated situation/simulated work task situation (e.g., Borlund, 2003a)

uine information needs–both when pilot testing and when carrying out the actual evaluation. This means that the test participants should prepare real, personal information needs which they search as part of the evaluation. Hence, genuine information needs function as a baseline against the simulated information needs, hereby acting as a control on the search interaction derived from the searching of simulated work task situations. In addition, the genuine information needs provide information about the systems' effect on real information needs. The inclusion of genuine information needs is also useful in the pilot test (requirement no. 5) because personal information needs can inspire 'realistic' and user-adaptable simulated work task situations.

The fourth requirement advises to permute the order of search jobs between the test participants so that no test participants are presented with the same simulated work tasks and their own personal information need in the same order. This is to neutralise any effect on the results in terms of bias of search interaction and relevance assessment behaviour of the test participants as well as the test participants' increasing system knowledge and possible knowledge of domain topicality of the simulated work tasks situations.

The fifth requirement concerns the ever-good test practice of pilot testing prior to actual evaluation. When pilot testing the test setting, the test procedure, the collected data, and the test participants' perceptions of the simulated work task situations are evaluated, and adjusted accordingly if required. As mentioned it is most useful to instruct the pilot-test participants to contribute with real, personal information needs as these needs can inspire simulated work task situations that are ideal for the group of test participants. If that is the case, then subsequent pilot testing is required in order to evaluate the test participants' view of the new simulated work task situation(s).

The sixth and final requirement points out the fact that the employed simulated work task situations must be depicted when the study is reported. Otherwise the reader is not able to assess the realism of the employed tailored simulated work task situation with respect to the target group of test participants, and hence assess the value and strength of the reported results of the study.

Simulated work task situations satisfy the experimental demands illustrated with the relevance and the cognitive revolutions (Robertson & Hancock-Beaulieu, 1992). These demands are that relevance should be judged in relation to the need rather than the request, and that an information need should be acknowledged as individual and dynamic, the process of the need formation being a situation-driven phenomenon. By applying simulated work task situations the test participants can work with personal and individual information need interpretations, which can develop and mature over session time for the same test participant. The dynamic nature of the information need formation is strongly connected to the dynamic nature of assessing relevance. By the application of simulated work task situations, non-binary relevance and the type of the information searching and retrieval processes involved in the use of an IIR system is available for studying. Schamber, Eisenberg and Nilan (1990, p. 774) draw the following conclusions on the nature of relevance and its role in information behaviour:

1. Relevance is a multidimensional cognitive concept whose meaning is largely dependent on users' perceptions of information and their own information need situations;
2. Relevance is a dynamic concept that depends on users' judgements of quality of the relationship between information and information need at a certain point in time; *and*
3. Relevance is a complex but systematic and measurable concept if approached conceptually and operationally from the user's perspective.

The first and second conclusions support both the application of situational relevance and non-binary relevance assessment. In addition, the third conclusion supports the application of simulated work task situations by implying that, while the number of relevant information objects retrieved is still a parameter to be measured, the relevance of an information object is defined not solely by the topic of the user's query, but by how useful the information contained in the retrieved information object is in relation to the information need and the underlying situation–as in real life.

6. CONCLUDING REMARKS AND FURTHER READINGS

The ambition of this paper has been to introduce IIR, and in particular IIR evaluation, because of the large quantity of IR research in IR evaluation and methodological issues (Järvelin, 2011). The point of departure has been historical, starting out with an introduction to the Cranfield model, to which IIR evaluation is the counterpart. The ASTIA and Cranfield tests constitute the empirical tradition of IR, and the experiences earned from these tests explain how IR ended up with the rigid definitions of the nature of the information need and relevance. The co-existing user-oriented IR research is exemplified with The MEDLARS test, and the development of end-user IR systems (OPACs) emphasised the need for alternative approaches to IR evaluation. In this paper illustrated with the Book House system for fiction retrieval and the research project of Okapi. The need for alternative approaches to IR evaluation is nicely summarised by Robertson and Hancock-Beaulieu (1992) in the form of the three revolutions: the cognitive, the relevance, and the interactive revolutions. The three revolutions present the modern and contemporary requirements to IIR evaluation. As an example of current IIR evaluation methodology the IIR evaluation model by Borlund is introduced (e.g., Borlund, 2000a; 2000b; 2003a). The IIR evaluation model meets the requirements of the three revolutions, not the least due to the employment of the test instrument of a simulated work task situation. Despite the qualities of the IIR evaluation model more research is needed. For example, a reoccurring issue is that of generalization of IIR evaluation results. This issue is also addressed by Belkin in his 2008 ECIR Keynote, when he points out how the contradictions between the necessity for realism and the desire for comparability and generalization have not yet been solved (Belkin, 2008, p. 52). Belkin is not the only one to comment on the need for further research on IIR evaluation. Järvelin (2011, p. 137) puts it as follows: "Information retrieval evaluation will not be remembered in history books for solving easy problems. Solving the difficult ones matters. Task-based and user-oriented evaluations offer such problems. Solving them can potentially lead to significant progress in the domain."

For further reading on IR systems evaluation from a historical point of view the reader is directed to the book in memory of Cleverdon edited by Spark Jones (1981) titled *Information Retrieval Experiment*. For more concrete guidelines the reader is recommended the hands-on paper by Tague-Sutcliffe (1992) with ten decisions to make when conducting empirical IR research, or the compendium paper by Kelly (2009) on methods for evaluating IIR systems with users. Also the ARIST chapters by Harter and Hert (1997), and Wang (2001) are recommendable when considering approaches, issues, and methods for IR systems evaluation and evaluation of information user behaviour. The book by Ingwersen and Järvelin (2005), which aims at integrating research in information seeking and IR, deserves attention, too. So does the recent book by Fidel (2012) that carefully outlines the framework for cognitive domain analyses deriving from the Book House project. The book by Xie (2008) on IIR in digital environments is also worth mentioning, and so is the excellent ARIST chapter by Ruthven (2008) which introduces IIR from the perspective of searching and retrieval. Finally, attention should be given the book edited by Ruthven and Kelly (2011) about interactive information seeking, behaviour, and retrieval.

ACKNOWLEDGEMENT

In memory of Professor Emeritus F. W. Lancaster, who passed away during the writing of this paper. He showed us the importance of thorough test design for the evaluation and study of systems performance and human interaction with information.

REFERENCES

Aitchison, J. & Cleverdon, C. (1963). *Aslib Cranfield research project: Report on the test of the Index of Metallurgical Literature of Western Reserve University*. Cranfield: The College of Aeronautics.

Beaulieu, M. & Jones, S. (1998). Interactive searching and interface issues in the Okapi Best Match Probabilistic Retrieval System. *Interacting with Computers*, 10, 237-248.

Beaulieu, M. (1997). Experiments on interfaces to support query expansion. *Journal of Documentation*, (53)1, 8-19.

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. *The Canadian Journal of Information Science*, (5), 133-143.

Belkin, N.J. (2008). Some(what) grand challenges for information retrieval. *ACM SIGIR Forum*, 42 (1), 47-54.

Belkin, N.J., Cool, C., Croft, W.B. & Callan, J.P. (1993). The effect of multiple query representation on information retrieval system performance. In R. Korfhage, E. Rasmussen, & P. Willett (Eds.), *Proceedings of the 16th ACM Sigir Conference on Research and Development of Information Retrieval. Pittsburgh*, 1993. New York: ACM Press, 339-346.

Belkin, N.J., Oddy, R. & Brooks, H. (1982). ASK for information retrieval: Part I. Background and theory. *Journal of Documentation*, 38 (2), 61-71.

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 B.W. Croft, A. Moffat, C.J. van Rijsbergen, R. Wilkinson, & J. Zobel (Eds.), *Proceedings of the 21st ACM Sigir Conference on Research and Development of Information Retrieval*. Melbourne, 1998. Australia: ACM Press/York Press, 324-331.

Borlund, P. (2000a). *Evaluation of interactive information retrieval systems*. Åbo: Åbo Akademi University Press. Doctoral Thesis, Åbo Akademi University.

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

Borlund, P. (2003a). The IIR evaluation model: A framework for evaluation of interactive information retrieval systems. *Information Research*, 8 (3). Retrieved from http://informationr.net/ir/8-3/paper152.html

Borlund, P. (2003b). The concept of relevance in IR. *Journal of the American Society for Information Science and Technology*, 54 (10), 913-925.

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, 142-148.

Cleverdon, C.W. & Keen, E.M. (1966). *Aslib Cranfield Research Project: Factors determining the performance of indexing systems*. Vol. 2: Results. Cranfield.

Cleverdon, C.W. (1960). *Aslib Cranfield Research Project: Report on the first stage of an investigation into the comparative efficiency of indexing systems*. Cranfield: the College of Aeronautics.

Cleverdon, C.W. (1962). *Aslib Cranfield Research Project: Report on the testing and analysis of an investigation into the comparative efficiency of indexing systems*. Cranfield.

Cleverdon, C.W., Mills, J. & Keen, E.M. (1966). *Aslib Cranfield Research Project: Factors determining the performance of indexing systems*. Vol. 1: Design.

Cool, C. & Belkin, N.J. (2011). Interactive information retrieval: History and background. In I. Rutven & D. Kelly (Eds.), *Interactive information seeking, behaviour and retrieval*. London: Facet Publishing, 1-14.

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

Ellis, D. (1996). *Progress and problems in information retrieval*. London: Library Association Publishing.

Fidel, R. (2012). *Human information interaction: An ecological approach to information behavior*. Cambridge, MA: MIT.

Goodstein, L.P. & Pejtersen, A.M. (1989). *The Book House: System functionality and evaluation*. Roskilde, Denmark: Risø National Laboratory, (Risø-M-2793).

Gull, C.D. (1956). Seven years of work on the organization of materials in the special library. *American Documentation*, 7, 320-329.

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

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.

Ingwersen, P. & Järvelin, K. (2005). *The turn: Integration of information seeking retrieval in context*. Dordrecht, Netherlands: Springer Verlag.

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

Järvelin, K. & Kekäläinen, J. (2000). IR evaluation methods for retrieving highly relevant documents. In N.J. Belkin, P. Ingwersen, & M.-K. Leong (Eds.), *Proceedings of the 23rd ACM Sigir Conference on Research and Development of Information Retrieval*. Athens, Greece, 2000. New York, N.Y.: ACM Press, 2000, 41-48.

Järvelin, K. (2011). Evaluation. In I. Rutven & D. Kelly (Eds.), *Interactive information seeking, behaviour and retrieval*. London: Facet Publishing, 113-138.

Kelly, D. (2009). Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends in Information Retrieval*, 3 (1-2), 1-224.

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, 119-142.

Lu, W., Robertson, S.E. & Macfarlane, A. (2007). CISR at INEX 2006. In N. Fuhr, M. Lalmas, and A. Trotman (Eds.), *Comparative Evaluation of XML Information Retrieval Systems: 5th International Workshop of the Initiative for the Evaluation of XML Retrieva (INEX 2006)*, Dagstuhl, Germany, LNCS 4518, Springer-Verlag, (2007), 57-63.

Lu, W., Robertson, S.E. & Macfarlane, A. (2006). Field-Weighted XML retrieval based on BM25. In N. Fuhr, M. Lalmas, S. Malik, & G. Kazai (Eds.), *Advances in XML Information Retrieval and Evaluation: Fourth Workshop of the INitiative for the Evaluation of XML Retrieval (INEX 2005)*, Dagstuhl, 28-30 November 2005, Lecture Notes in Computer Science, Vol 3977, Springer-Verlag,

Marchionini, G. (2006). Toward human-computer information retrieval bulletin. In June/July 2006 *Bulletin of the American Society for Information Science*. Retrieved from http://www.asis.org/Bulletin/Jun-06/marchionini.html

Martyn, J. & Lancaster, F.W. (1981). *Investigative methods in library and information science: An introduction*. Virginia: Information Resources Press. 1981. (2nd impression September 1991).

Pejtersen, A.M. & Austin, J. (1983). Fiction retrieval: Experimental design and evaluation of a search system based on users' value criteria (Part 1). *Journal of Documentation*, 39 (4), 230-246.

Pejtersen, A.M. & Austin, J. (1984). Fiction retrieval: Experimental design and evaluation of a search system based on users' value criteria (Part 2). *Journal of Documentation,* 40 (1), 25-35.

Pejtersen, A.M. & Fidel, R. (1998). A framework for work centered evaluation and design: A case study of IR on the web. Grenoble, March 1998. [Working paper for MIRA Workshop, Unpublished].

Pejtersen, A.M. & Rasmussen, J. (1998). Effectiveness testing of complex systems. In M. Helander (Ed.), *Handbook of human-computer interaction*. Amsterdam: North-Holland, 1514-1542.

Pejtersen, A.M. (1980). Design of a classification scheme for fiction based on an analysis of actual user-librarian communication and use of the scheme for control of librarians' search strategies. In O. Harbo, & L. Kajberg (Eds.), *Theory and application of information research. Proceedings of the 2nd International Research forum on Information Science*. London: Mansell, 146-159.

Pejtersen, A.M. (1989). A library system for information retrieval based on a cognitive task analysis and supported by an icon-based interface. In *Proceedings of the 12th Annual International ACM SIGR Conference on Research and Development in Information Retrieval* (SIGIR 1989), ACM, 40-47.

Pejtersen, A.M. (1991). *Interfaces based on associative semantics for browsing in information retrieval*. Roskilde, Denmark: Risø National Laboratory, (Risø-M-2883).

Pejtersen, A.M. (1992). New model for multimedia interfaces to online public access catalogues. *The Electronic Library*, 10 (6), 359-366.

Pejtersen, A.M., Olsen, S.E. & Zunde, P. (1987). Development of a term association interface for browsing bibliographic data bases based on end users' word associations. In I. Wormell (Ed.), *Knowledge engineering: expert systems and information retrieval*. London: Taylor Graham, 92-112.

Rasmussen, J., Pejtersen, A.M. & Goodstein, L.P. (1994). *Cognitive systems engineering*. N.Y.: John Wiley & Sons.

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

Robertson, S.E. (1981). The methodology of information retrieval experiment. In K. Sparck Jones (Ed.), *Information retrieval experiments*. London: Buttersworths, 9-31.

Robertson, S.E. (1997a). Overview of the Okapi Projects. *Journal of Documentation*, 53 (1), 3-7.

Robertson, S.E. (Ed.). (1997b). Special issue on Okapi. *Journal of Documentation*, 53 (1).

Robertson, S.E., Lu, W. & MacFarlane, A. (2006). XML-structured documents: Retrievable units and inheritance. In H. Legind Larsen, G. Pasi, D. Ortiz-Arroyo, T. Andreasen, & H. Christiansen (Eds.), *Proceedings of Flexible Query Answering Systems 7th International Conference*, FQAS 2006, Milan, Italy, June 7-10, 2006, LNCS, 4027, Springer-Verlag, (2006), 121-132.

Robertson, S.E., Walker, S. & Beaulieu, M. (1997). Laboratory experiments with Okapi: Participation in the TREC programme. *Journal of Documentation*, 53 (1), 20-34.

Ruthven, I. & Kelly, D. (Eds.). (2011). *Interactive information seeking, behaviour and retrieval*. London: Facet Publishing.

Ruthven, I. (2008). Interactive information retrieval. *Annual Review of Information Science and Technology*, 24 (1), 2008, 43-91.

Salton, G. (1972). A New comparison between conventional indexing (MEDLARS) and automatic text processing (SMART). *Journal of the American Society for Information Science*, (March-April), 75-84.

Salton, G. (1981). The smart environment for retrieval system evaluation: Advantages and problem areas. In K. Sparck Jones (Ed.), *Information retrieval experiments*. London: Buttersworths, 316-329.

Sanderson, M. (2010). Test collection based evaluation of information retrieval systems. *Foundations and Trends in Information Retrieval*, 4 (4), 247-375.

Saracevic, T. (1995). Evaluation of evaluation in information retrieval. In E.A Fox, P. Ingwersen, & R. Fidel (Eds.), *Proceedings of the 18th ACM Sigir Conference on Research and Development of Information Retrieval*. Seattle, 1995. N.Y.: ACM Press, 138-146.

Schamber, L. (1994). Relevance and information behavior. In M.E. Williams (Ed.), *Annual Review of Information Science and Technology (ARIST)*. Medford, NJ: Learned Information, INC., 29, 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), 755-775.

Sharp, J. (1964). Review of the Cranfield-WRU test literature. *Journal of Documentation*, 20 (3), 170-174.

Sparck Jones, K. (1971). *Automatic keyword classification for information retrieval*. London: Buttersworths.

Sparck Jones, K. (1981a). Retrieval system tests 1958-1978. In K. Sparck Jones (Ed.), *Information retrieval experiments*. London: Buttersworths, 213-255.

Sparck Jones, K. (1981b). The Cranfield tests. In K. Sparck Jones (Ed.), *Information retrieval experiments*. London: Buttersworths, 256-284.

Sparck Jones, K. (Ed.). (1981c). *Information retrieval experiments*. London: Buttersworths.

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.

Swanson, D.R. (1965). The evidence underlying the Cranfield results. *Library Quarterly*, 35, 1-20.

Swanson, D.R. (1986). Subjective versus objective relevance in bibliographic retrieval systems. *Library quarterly*, 56, 389-398.

Tague-Sutcliffe, J. (1992). The pragmatics of information retrieval experimentation, revisited. *Information Processing & Management*, 28(4), 467-490.

Thorne, R.G. (1955). The efficiency of subject catalogues and the cost of information searches. *Journal of Documentation*, 11(3), 130-148.

Voorhees, E.M. & Harman, D.K. (2005a). The text retrieval conference. In E.M. Voorhees & D.K. Harman (Eds.). *TREC: Experiment and evaluation in information retrieval*. Cambridge, Massachusetts: The MIT Press. 3-19.

Voorhees, E.M. & Harman, D.K. (Eds.) (2005b). *TREC: Experiment and evaluation in information retrieval*. Cambridge, Massachusetts: The MIT Press.

Walker, S. & De Vere, R. (1990). *Improving subject retrieval in online catalogues: 2. Relevance feedback and query expansion*. London: British Library. (British Library Research Paper 72).

Walker, S. (1989). The Okapi online catalogue research projects. In *The Online catalogue: developments and directions*. London: The Library Association, 84-106.

Wang, P. (2001). Methodologies and methods for user behavioral research. In M.E. Williams (Ed.), *Annual Review of Information Science and Technology*, 34, 1999, 53-99.

Wilson, M. (2011). Interfaces for information retrieval. In I. Rutven, & D. Kelly (Eds.), *Interactive information seeking, behaviour and retrieval*. London: Facet Publishing, 139-170.

Xie, I. (2008). *Interactive information retrieval in digital environments*. IGI Publishing.