Advanced Topics in Information
Retrieval, Springer 2011: 91-118
Ed.: M. Melucci & R. Baetza Yeates

Chapter 4
The User in Interactive Information Retrieval Evaluation

Peter Ingwersen

Abstract This chapter initially defines what characterizes and distinguishes research frameworks from research models. The Laboratory Research Framework for IR illustrates the case. We define briefly what is meant by the concept of research design, including research questions, and what this chapter regards as central IIR evaluation research settings and variables. This is followed by a description of IIR components, pointing to the elements of the Integrated Cognitive Research Framework for IR that incorporates the Laboratory Framework in a contextual manner. The following sections describe and exemplify: (1) Request types, test persons, task-based simulations of search situations and relevance or performance measures in IIR; (2) Ultra-Light Interactive IR experiments; (3) Interactive-Light IR studies; and (4) Naturalistic field investigations of IIR. The chapter concludes with a summary section, a reference list and a thematically classified bibliography.

4.1 Introduction

Since the dawn of Information Retrieval (IR) experimentation and IR evaluation two approaches have been predominant. The mainstream laboratory-based IR research and evaluation framework, also named the Cranfield Model or Laboratory IR Research Framework (Baeza-Yates and Ribeiro-Neto, 1999; Belew, 2000; Ingwersen and Järvelin, 2005), and the user-oriented perspectives on interactive IR (IIR) (Belkin and Vickery, 1985; Järvelin, 2007). The latter matured somewhat later, are commonly not adhering to one single re-

Peter Ingwersen
Royal School of Library and Information Science, Denmark; Oslo University College, Norway
Birketinget 6, DK 2300 Copenhagen S, Denmark, http://www.iva.dk/pi e-mail: pi@iva.dk

search framework, but display variations between models and methodologies (Ingwersen and Järvelin, 2005).

This chapter puts forward an integrated and contextual perspective on IR experimentation and evaluation, founded upon a cognitive approach to IR (Ingwersen and Järvelin, 2005), as an alternative to the Laboratory IR Research Framework. Hence, this perspective is regarded an attempt to create an Integrated Cognitive Research Framework that may cover a variety of interactive IR models as an umbrella and provide a range of workable research methodologies. The framework integrates human actors, like searchers or authors of information, with their socio-cultural-organizational *and* systemic contexts. The motivation behind is twofold: 1) it is not sufficient to postulate an alternative epistemological perspective to IR (viz. the integrated cognitive approach) *without also* providing the consequential research design tools and methodologies; 2) in recent years laboratory researchers have increasingly asked for the provision of such tools and methodologies, so that they might address user-IR system interaction from a more contextual perspective.

This chapter initially defines what we believe characterizes and distinguishes research frameworks from research models. The Laboratory IR Research Framework illustrates the case. We define briefly what is meant by the concept of research design, including research questions, and what this chapter regards as central IIR evaluation research settings and variables. This is followed by a description of IIR components, pointing to the central elements of the Integrated Cognitive Research Framework for IR that incorporates the Laboratory Research Framework in a contextual manner.

The following sections describe and exemplify 1) Request types, test persons, task-based simulations of search situations and relevance or performance measures in IIR; 2) Ultra-light Interactive IR experiments; 3) Interactive-Light IR studies; and 4) Naturalistic field investigations of IIR. The chapter concludes with a summary section and a double-purpose reference section. References are listed as in the text. Then follows a more comprehensive bibliography categorized according to the structure of the chapter, including many additional bibliographic entries.

4.2 Research Frameworks, Models and other Central Concepts

A *research framework* for IR describes and models the central objects to study, their relationships as well as changes in objects and in their relationships that (may) affect the functioning of the IR system and interactive processes. Further, it outlines promising goals and methods of research (Ingwersen and Järvelin, 2005, 11-13). Research frameworks typically contain (tacit) knowledge and shared assumptions on its ontological, conceptual, factual, epistemological and methodological elements. *Models* are precise (often

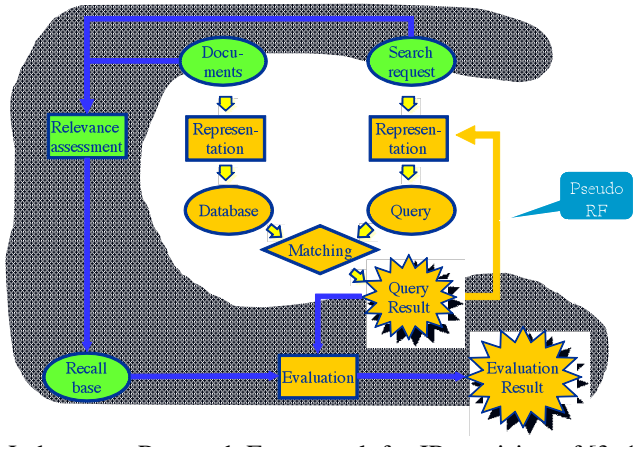

Fig. 4.1: The Laboratory Research Framework for IR; revision of (Ingwersen and Järvelin, 2005, 115).

formal) representations of objects and relationships or processes within a research framework.

Examples of formal models in IR are probabilistic, vector space, language, logical, quantum-theoretical, etc. models that compete under the umbrella of the Laboratory Research Framework for IR. IR Models may indeed also be graphic and in principle encompass human actors and organizations. Fig. 4.1 depicts graphically the Laboratory Research Framework for IR with the generalized Cranfield model at its centre.

The framework displays its central variables (objects located in the laboratory cave as seen in a vertical cut, with the entry of the cave to the right), relationships and processes to be carried out or/and studied. It holds at its centre the *Cranfield Model of IR* containing a set of requests (topics in TREC) represented (indexed) as queries, a collection of documents, their representation (indexing) in a database (nowadays including the full documents), and a matching algorithm. Obviously, if the same sets of queries and documents are applied for all experiments, whereby they are controlled or variations are statistically neutralized, one may simply stick to one indexing algorithm during experimentation (then also controlled) and solely vary the matching algorithms. Or do the opposite. This is the robust natural science-like research design philosophy behind the Laboratory Research Framework for IR that makes it so successful: only one independent variable is in play at a time in each experiment.

The Laboratory Research Framework in addition contains the process of *relevance assessment* made by human assessors, one for each request/topic, commonly based on a pooling principle of the retrieved documents per query.

For the total set of requests/topics the relevant documents are stored in a *recall base* and compared to the *query results* for each retrieval run in the experiment. The *evaluation result* can then be calculated in terms of performance measures of various kinds (Baeza-Yates and Ribeiro-Neto, 1999; Belew, 2000). Commonly, the relevance assessment scale is binary. In addition, the Laboratory Research Framework allows for pseudo relevance feedback (pseudo RF) to be studied. Pseudo RF works as a *simulation* of RF behaviour made by a human searcher (Magennis and van Rijsbergen, 1997; White, 2006; White et al, 2005b). The process of *relevance assessment* constitutes the only weak element of the framework. This is because the human assessor is – *human* – i.e., he/she would become saturated when judging documents after a while or might become sloppy, non-motivated, bored, etc.; however, it is argued that since these characteristics are equal for all assessors over all requests/topics, they are also equally distributed across the involved laboratories with their competing algorithms. With enough requests/topics the variation of relevance judgments becomes statistically neutralized, see e.g. studies by Voorhees (1998) over the last decade.

4.2.1 Research Design and IIR Research Setting Types

The reason for detailing the Laboratory IR Research Framework is that it constitutes a central element of the Integrated Cognitive Research Framework for IR. Essentially, the latter framework pushes the experimental situation outside the laboratory cave into the context of reality. The Integrated Research Framework for IR is called 'cognitive' because it adheres to the epistemological cognitive viewpoint, in which the contextual (social and systemic) elements of an actor influence and become influenced by that actor during interaction. IR systems are seen as systems supporting human cognition (Ingwersen and Järvelin, 2005, 23-31). The problem when setting up a laboratory experiment that involves users is to keep the experimental situation under *control* and, at the same time, allow the test persons to have *cognitive freedom*. Whilst control relates to something artificial and static, the latter associates to realism and situational dynamism.

A solid *research design* generates the research problem – *what* is to be studied? It consists of research questions – *why*? – and the decisions concerning the degree of human involvement in experiments – the *control* issue. Research design deals with setting up the data collection and deciding the methodological approach – *how* and *when*. This includes the range of research outcomes and thus the data analysis method. We should always remember that the aim of IR is twofold: designing and evaluating IR systems *and/or* understanding searcher behaviour in context. The Laboratory Research Framework is isolated to deal with the former whilst the Integrated Cognitive Research Framework seeks to circumscribe both aims.

Research questions should ideally be answered by the investigation at hand. They must be concise and meaningful statements of the research goal(s), i.e. the research outcomes. Hypotheses are closely related to the research questions by serving as their precursors or motivation. They form beliefs or predictions about relationships of objects that arises in accordance with the research framework, model(s) or theory. An example of a hypothesis could be '[knowledge of] multi-evidence of a searcher's information situation may improve retrieval performance through query expansion, compared to initial searcher request formulation and use of pseudo-relevance feedback' (Kelly and Fu, 2007).

In the remaining of this chapter the following types of IR *research settings* are used as investigations:

- Laboratory experiments: no test persons participate. Investigations deal with performance tests of kinds of algorithms, simulations of searcher behaviour or log analyses (not treated in this chapter).
- Laboratory study: test persons participate:
  - Ultra-light IR study: investigations of short-term IR interaction of 1-2 retrieval runs;
  - Light IR study: investigations of session-based multi-run IR interaction;
- Field experiment: experimental situation tested in natural setting with test persons;
- Field study: investigation of system performance or human behaviour in natural setting with test persons;
- Longitudinal studies

In IR investigations the common tradition is that experiments and studies are carried out in a rigorous statistical manner, i.e., a sufficient number of test persons and/or search jobs are applied to the setting in order to produce statistically valid results (within its data limits). This tradition adheres to the performance measurement history of the Cranfield investigations, dating back to the 1960s. Case studies, although accepted in most social science fields, are not really accepted in the Laboratory Research Framework, but acceptable in the Integrated Cognitive Research Framework, in particular concerning behavioural investigations of IR phenomena. We may here observe a link to Human-Computer Interaction (HCI), which by nature commonly do *not* operate with large groups of test persons or search jobs (Hornbaek, 2006).

Regardless the research setting and research frameworks applied to IR investigations the following variables are at play (Ingwersen and Järvelin, 2005; Fidel and Soergel, 1983):

- Independent variables: the cause, e.g., different IR models; interface functionality; searcher knowledge level;
- Dependent variables: the effect, e.g., as measured by some performance measure of retrieval result (MAP; DCG) or usability measures (search time; clicks);

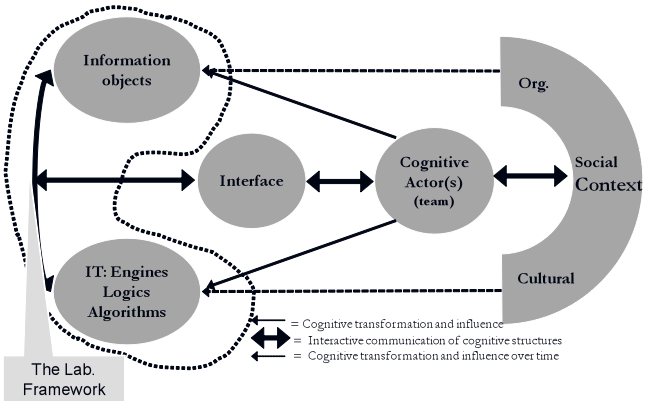

Fig. 4.2: Central components of interactive IR – the basis of the Integrated Cognitive Research Framework for IR. Revised version of (Ingwersen and Järvelin, 2005, p. 261)

- Controlled variables: variables held constant, statistically neutralized or randomized, e.g., document collection; retrieval model; assigned topics or simulated task situations (the search jobs); test persons;
- Hidden variables: moderating or intervening (may create bias), e.g., variations of searcher domain knowledge levels; de-motivation of test persons; no up-to-date collection.

As stated above in the Laboratory Research Framework the number of variables is limited and although each may take many values, the setting is fairly easy to control. In TREC this may be the case also because each TREC track feeds on its own document collection and applying tailored search jobs. The experimental situation would be more loose if the one and same collection was applied to a mixture of research goals (tracks), e.g., mixing different kinds of requests and document types. The chance of hidden variables and lack of control would definitively increase in such rather naturalistic scenarios, simulating a digital library or integrated searching.

4.2.2 Central IIR Components

Fig. 4.2 displays the six central components of IIR, with the Laboratory Research Framework (dotted figure) to the left, covering Information Objects, the IT component and the interaction them in between. The Interface, the Cognitive Actor and the Socio-cultural and Organizational context – and the remaining portion of the Interaction processes – are outside the frame-

work, and outside the cave stipulated in Fig. 4.1. The one-way arrows signify influence and transformation, e.g., from the communities of actors (the socio-cultural and organizational context) towards the documents or the IT components over time (dotted arrows), or the direct act of creation of objects or IT elements (straight arrows). The interaction processes consist of *IR Interaction* between Cognitive Actor and Interface component (request formulations and other statements from searcher) and further into the IT-Information-Object interaction, via query formulations. This is where the access to IR systems takes place. The interaction between actor and socio-cultural and organizational context constitutes *Social Interaction*.

4.2.3 The Integrated Cognitive Research Framework for IR

The six central components, Fig. 4.1, constitute the nine dimensions of the Integrated Cognitive Research Framework. Each dimension holds a number of variables, each with two or more values (Ingwersen and Järvelin, 2005, 2007):

- Natural work task variables, from the socio-cultural and organizational context;
- Natural search task variables, from the socio-cultural and organizational context;
- Actor characteristics variables, actor's personal characteristics;
- Perceived work task variables, actor's perception of natural work task;
- Perceived search task variables, actor's perception of natural search task;
- Document variables, dealing with all information object features and representations;
- Algorithmic search engine variables, concerned with features of the IT component;
- Algorithmic interface variables, dealing with interface functionalities;
- Access and interaction variables, concerning all features of IR and social interaction.

The work task is viewed as the underlying motivation for searchers to have an information need and the search task as the instrumental activity that may lead to solving the work task. Work tasks may be job-related or associated with non-job but daily-life situations (Ingwersen and Järvelin, 2005, 2007). Work and search tasks can be natural, i.e. really existing in the world – or they may become perceived and interpreted by actors. Manuals or Good Laboratory Practice (or like documentation) are examples of natural work tasks described in the real world. There exists of course a difference between such *natural* work or search tasks and the range of *assigned* ones IR research

makes use of. This range of assignments goes from semantically open simulated work task situations (cover stories) (Borlund, 2003b) over semantically closed situations to TREC topics with description and narrative or simply to one-line or two-term assigned requests.

The contents of the nine dimensions relies on empirical or analytic investigations carried out over the last three decades on IIR and laboratory IR. For instance, the variables and values constituting the Interface dimension derives from a mixture of the MONSTRAT and MEDIATOR models generated by Belkin et al (1983) and Ingwersen (1992), respectively.

The lists of variables forming up the Integrated Cognitive Research Framework originates from (Ingwersen and Järvelin, 2005, 356-357) and are discussed by Ingwersen and Järvelin (2007), in particular associated with relevance and interaction. Tables 4.1 and 4.2 display the original multidimensional array of dimensions and variables.

The framework is intended to operate with a maximum of three independent variables, each containing binary values. The variables must be treated in pairs and can be illustrated by the following typical IIR variables:

- Interface function X, value a/b – e.g. response generation (presentation form): Yahoo snippet vs. bibliographic record;
- IS&R knowledge – search expertise, having values none/much;
- Natural/assigned work task type – e.g. size: richly vs. poorly defined.

The array of dimensions, Tables 4.1, 4.2, is later used to mark up the specific variables involved in the three research design examples outlined below according to the Integrated Cognitive Research Framework for IR.

4.3 IR Interaction – Research Designs with Test Persons

Regardless how IR interaction laboratory studies or field experiments are developed the research designs must deal with request types used in the setting, number of test persons and search jobs involved, and the application of appropriate relevance, usability and evaluation measures. Section 4.3.1 discusses such issues of interactive research design. Sections 4.3.2, 4.3.3 and 4.3.4 outline, discuss and exemplify, respectively, IR interaction ultra-light, interactive-light and naturalistic research scenarios.

Natural Work Tasks (WT) & Org	**Natural Search Tasks (ST)**	**Actor**	**Perceived Work Tasks**	**Perceived Search Tasks**
WT Structure	ST Structure	Domain Knowledge	Perceived WT Structure	Perceived Information Need Content
WT Strategies & Practices	ST Strategies & Practices	IS&R Knowledge	Perceived WT Strategies & Practices	Perceived ST Structure / Type
WT Granularity, Size & Complexity	ST Granularity, Size & Complexity	Experience on Work Task	Perceived WT Granularity, Size & Complexity	Perceived ST Strategies & Practices
WT Dependencies	ST Dependencies	Experience on Search Task	Perceived WT Dependencies	Perceived ST Specificity & Complexity
WT Requirements	ST Requirements	Stage in Work Task Execution	Perceived WT Requirements	Perceived ST Dependencies
WT Domain & Context	ST Domain & Context	Perception of Socio-Org Context	Perceived WT Domain & Context	Perceived ST Stability
		Sources of Difficulty		Perceived ST Domain & Context
		Motivation & Emotional State		

Table 4.1: Five dimensions of variables of the Integrated Cognitive Research Framework (Ingwersen and Järvelin, 2005, p. 356)

4.3.1 Search Job Design, Simulated Task Situations, Test Persons and Evaluation Measures

Search job design

The choice of appropriate request types to form the search jobs in IR interaction is important and depends on the research questions. Originally the Laboratory Research Framework applied assigned and quite rich *topical* information need formulations (here regarded 'requests'), mainly owing to the best match retrieval models' preference of a substantial number of search keys in order to function well. This issue has been softened in recent years, predominantly due to the scarcity of search keys applied in natural Web searching; see e.g. the TREC developments (Harman, 1996).

Document and Source	**IR Engines IT Component**	**IR Interfaces**	**Access and Interaction**
Document Structure	Exact Match Models	Domain Model Attributes	Interaction Duration
Document Types	Best Match Models	System Model Features	Actors or Components
Document Genres	Degree of Document Structure and Content Used	User Model Features	Kind of Interaction and Access
Information Type in Document	Use of NLP to Document Indexing	System Model Adaption	Strategies and Tactics
Communication Function	Document Metadata Representation	User Model Building	Purpose of Human Communication
Temporal Aspects	Use of Weights in Document Indexing	Request Model Builder	Purpose of System Communication
Document Sign Language	Degree of Request Structure and Content Used	Retrieval Strategy	Interaction Mode
Layout and Style	Use of NLP to Request Indexing	Response Generation	Least Effort Factors
Document Isness	Request Metadata Representation	Feedback Generation	
Document Content	Use of Weights in Requests	Mapping ST History	
Contextual Hyperlink Structure		Explanation Features	
Human Source (see Actor)		Transformation of Messages	
		Scheduler	

Table 4.2: Four dimensions of variables of the Integrated Cognitive Research Framework (Ingwersen and Järvelin, 2005, p. 357).

First of all, in IIR the 'query' is the retrieval mechanism's translation of the 'request formulation', according to its logic (Ingwersen and Järvelin, 2005). In command-driven IR systems like Thomson-Reuter's Dialog online Service, Medline and all web retrieval engines, the searcher is responsible for this translation in advanced search mode. The central request types applied in interactive IR investigations belong to three sets of characteristics: 1) whether they are *natural* or *assigned*; 2) whether they are content-*rich* or *poor*; 3) depending on input features and outcome, i.e., whether they are *topical* (in the TREC sense), *factual*, *known-item* like, or concerned with other *metadata*.

Assigned requests may take the form as: a) *simplistic* request formulations that commonly are context-free, as in TREC 'topics', in which the title, description and narrative do not explain why the request is posed or exists –

the TREC 'topic' title alone may form the shortest or most simplistic assigned request type; b) '*query by example*' where an information object (a photo, a publication, a tune ...) function as request formulation and a goal may be to retrieve something similar to that (Campbell, 2000) – commonly, such request types are quite contextual; c) *simulated work task situation* or cover story (Borlund, 2003b) – see below. By assigning requests a), b) c) the researcher attempts to keep the investigation under control – in contrast to allowing natural information need situations to occur in the experiments or study, which entail less control of the research situation.

Natural information need situations should be applied to the adequate context, i.e., the document collection characteristics and themes must be known to the test persons in order for them to generate appropriate requests for information. One should note that it is quite difficult to make searchers generate more that one-two different natural information needs each (i.e. per week or so in the same document environment). If forced, the same person often produces several information needs that look alike or are facets of the same core. Obviously, the researcher does not really know the retrieval outcome of natural requests before they have been searched in the given system.

Simulated search jobs

Simulated work task situations (Borlund, 2003b) consist of a 'story' that explains the search job the test person is supposed to do and why, e.g. a description of a job-related work task or a daily-life associated task situation, see Fig. 4.3. The idea is that when given to the test person the story provides the underlying reason or context for potential *interpretations* made by the person, thus posing an information request to the system. By giving the same simulated situation to all the test persons the research design is to a certain extent controlled. If the situation or cover story is content-rich and specific the potential for interpretation is more limited (increased control) than if the story is formulated in general and few terms. This issue deals with degree of *semantic openness* – Case of Fig. 4.3a is more closed than Case of Fig. 4.3b.

The Borlund evaluation package for IIR (Borlund, 2003b) contains recommendations as to data collection, in particular involving simulated work task situations, their styles and designs, and performance analysis applying alternative measures of performance. An essential feature of such situations is the degree of motivation they provide the test persons. They must be tailored to realistic and motivating scenarios in order to be carried out successfully. Because of their nature the subjects' perceptions and interpretations tend to promote targeted searching behaviour (for facts, topical and known item retrieval) rather than exploratory searching behaviour. One may profit from mixing simulated works task situations with natural ones, as shown by Kelly et al (2005); see also below, Section 3.5.

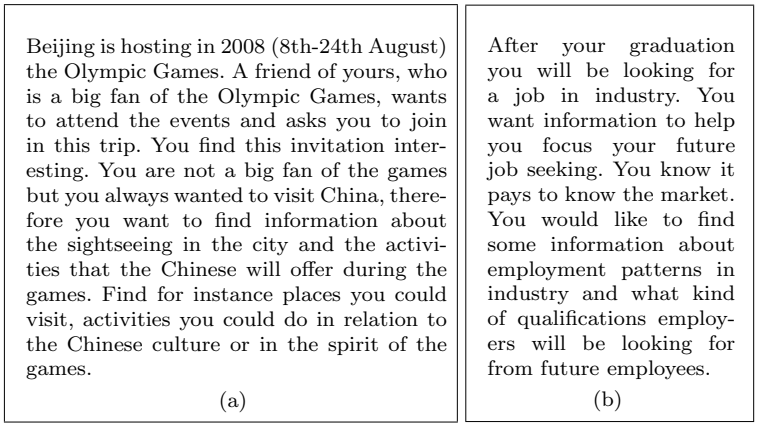

Fig. 4.3: Two examples of simulated work task situations.

Number of test persons and search jobs

The number depends on the research questions; but foremost it depends on the number of variables involved in the investigation. There are some rules of thumb to be applied (Ingwersen and Järvelin, 2005, 367). The central point is always to have at least 30 analysis entities in each cell of the result matrix.

If the study concerns *behavioural issues* many test persons are required in order to control potential human variation. At least 30 persons would be preferable, performing 2-3 search jobs each, if we are dealing with one variable with two values or two single value variables. Additional variables or values would entail additional search jobs per person.

If one investigates *retrieval performance*, involving test persons in IR interaction studies as done below, many search jobs per person, assigned or natural, are required but the number of test persons may be less than 30. The issue here is to control search job variations; but at the same time the knowledge characteristics of the test persons must be known and preferably similar across the subjects. Otherwise the end result can be biased because some persons stick out from the average – turning into hidden variables instead of being controlled or neutralized. 30 entities are still the magic number in the result matrix, in order to be statistically significant. As an example one might deal with two independent variables in a study: two different groups of test persons (say medical doctors and nurses) vs. two different retrieval models (probabilistic and PageRank). The result matrix holds four cells, each with 30 analysis entities, thus providing 120 entities in total. To do this research design with 10 medical doctors and 10 nurses as test persons $2 \times 3 = 6$ search jobs per person are required. 5 nurses will carry out search jobs 1-3 on the probabilistic machine and jobs 4-6 on the PageRank machine; the

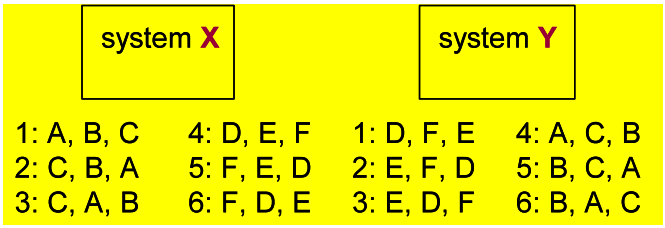

Fig. 4.4: Latin Square design with test persons (1-6) and search jobs (A-F) investigating two systems (X and Y) (Ingwersen and Järvelin, 2005, pp 253-254)

other 5 nurses do the search jobs in opposite order, see Fig. 4.4 for typical Latin Square design illustration. The same execution pattern is applied to the medical doctors and the 2×3 search jobs. In addition the search job sequence is permuted; see Fig. 4.4. No repetition has taken place (Ingwersen and Järvelin, 2005, 253-254) and each analysis cell in the matrix will hold 30 events.

This design is workable and statistically valid but may not satisfy the TREC research scenarios because of TREC's *competitive* nature. With 30 analysis entities only, the final rank order of the competing systems will not be stable. In TREC terms at least 50 entities must be present in each analysis cell in order to satisfy the competition principle (Voorhees, 1998); indeed it has been shown that if performance measures are done by MAP at quite *shallow ranking levels*, e.g., at the realistic ranks of top-10 or top-15 documents, more than 60 search jobs are required to maintain result sequence stability (Sanderson and Zobel, 2005). In the medical case above that implies that either a) each test person must perform 12 search jobs, or b) the number of test persons must be doubled to 2×20 subjects. The former is doable over a series of search sessions whilst the latter alternative often is cumbersome to fulfil owing to difficulty in getting enough persons with similar training and knowledge levels.

Performance measures

In IIR one may apply the traditional performance measures like precision, precision @ n, recall, Mean Average Precision (MAP), etc. They are commonly based on *binary* relevance assessments. Alternatively and more realistically *graded relevance* measures can be applied, as proposed, tested and generalized by Kekäläinen (2005) and Kekäläinen and Järvelin (2002b). Similarly, novel measures are applicable, like the DCG family of performance indicators that observe the degree of success by the retrieval engine of pushing up the

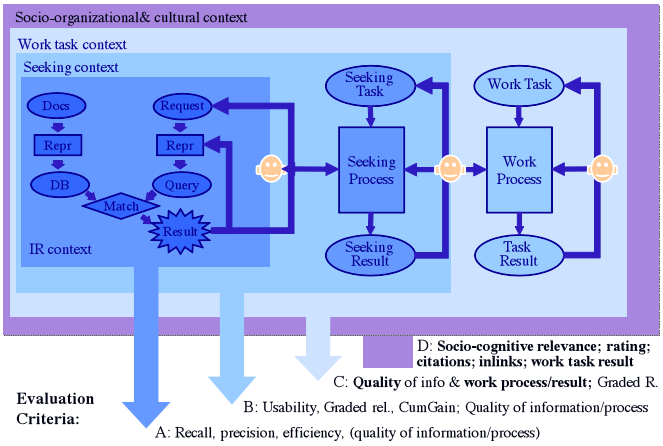

Fig. 4.5: The Integrated Cognitive Research Framework; relevance criteria, revision of (Ingwersen and Järvelin, 2005, p. 322).

relevant documents on the retrieval ranking compared to an ideal ranking sequence (Järvelin and Kekäläinen, 2002). The graded relevance issue and the degree of liberal assessment of relevance made in TREC scenarios have been tested by Sormunen (Sormunen, 2002).

Fig. 4.5 demonstrates the Integrated Cognitive Research Framework for IR with the Laboratory Research Framework to the left and the extension into an increasing degree of context from centre to right-hand side. In this model Laboratory IR is regarded in context of information seeking activities and work task processes. Simultaneous with the contextualization different novel performance indicators and other relevance/usefulness measures come into play. For instance, *usability* measures (Hornbaek, 2006) and the DCG family appear in association with IR interaction ultra-light and light investigations that in principle involve information seeking activities. Associated with usability measures, such as: display time; hovering over objects; amount of views and clicks; number of objects assessed; selection patterns; perception of ease; satisfaction; would be relevant dependent variables in interactive IR investigations.

Moving into the work task activity realm or even into the sociail and organizational context, i.e., into natural field experiments or studies, the *work task result* as well as several *socio-cognitive* (Cosijn and Ingwersen, 2000) or *social utility indicators* become useful measures of retrieval and system performance: peer reviewing results (e.g. in conference reviewing scores); density of social tagging; rating; citation counts; inlink volume; visits, search and download events; work task result; etc. Which measures to apply depend on

the actual research questions and the nature of the independent variables in the actual research design.

4.3.2 IR Interaction Ultra-Light Studies

Ultra-light IR studies are *laboratory* investigations of short-term IR interaction that consists of 1-2 retrieval runs with participation of test persons. The motivation behind this quite restricted form of IR interaction is to test results made from laboratory simulations, for instance of relevance feedback (Magennis and van Rijsbergen, 1997; White, 2006; White et al, 2005b), or to test new hypotheses based on other research, e.g., from information seeking studies, in a highly controlled environment.

The advantage of ultra-light interactive IR is that the researcher has two alternative research design approaches, owing to the little iteration during man-machine interaction: 1) applying assigned search jobs, either as TREC-like topics or in the form of simulated work task situations explained above; or 2) applying real-life natural search jobs generated by the test persons themselves.

In the first case, the *existing* recall base (Table 4.1), holding documents already assessed for relevance for each search job, can be re-used for performance measurements; the participating test persons may in addition supply novel relevance assessments after the first run, which may be compared to the existing ones in the test collection or compared to pseudo-RF. Existing test collections may hence be applied in this research design with the inclusion of a quite substantial number of relevant documents for performance measurements. The reason behind the re-use potential of the relevance base is that the test persons only *once* observe the ranked documents, just like the original assessor. For example, the first retrieval run is carried out by the searcher (or done automatically), followed by a relevance feedback (RF) process done by the person, followed by the second run from which the results can be compared to the recall base, see Fig. 4.6. The test person has suffered from *similar learning effects* as the original test collection assessor. However, if ensuing human RF activities and retrieval runs should be made, learning effects *will* surface and the research scenario makes the original assessments totally out of tune with the current test person's relevance perception. The research design then moves into an IR Interaction-light scenario.

The second alternative does not allow for re-use of the recall base in a test collection. With natural information needs/task situations all relevance assessments must be made by the test persons themselves from scratch. This research design is more realistic, although still containing a maximum of two runs under the ultra-light label. The advantage is that collections tailored to the research questions may be applied in the research design. Graded

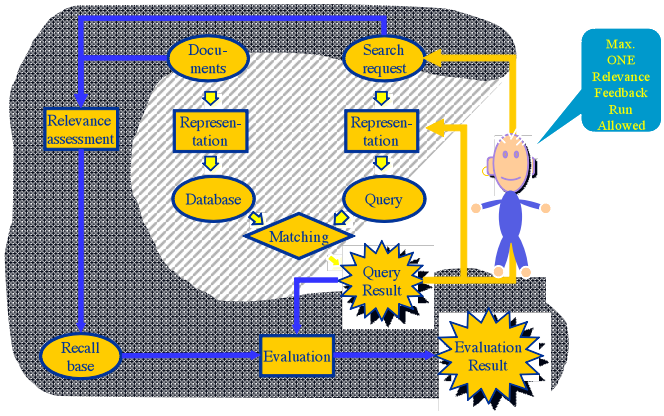

Fig. 4.6: IR interaction Ultra-light – short-term IR – revision of (Ingwersen and Järvelin, 2005, p. 5).

relevance can be applied by the test persons (Kekäläinen, 2005; Kekäläinen and Järvelin, 2002b).

In both research designs several pseudo RF runs may be applied prior to the single human RF run, thus allowing for more elaborate automatic / algorithmic experiments, but still involving searchers.

The disadvantages of IR interaction Ultra-light studies are: 1) the research design is *limited in realism* with only one run with human perception and interpretation involved; context features are hardly at play – interestingly, this scenario corresponds to that of the probabilistic retrieval model (Ingwersen and Järvelin, 2007) in order to get the model to function properly; 2) the second alternative with natural search jobs replacing assigned ones may only produce a *small number of assessments* owing to *assessor saturation* (Ingwersen and Järvelin, 2005; Borlund, 2003b). This saturation facet of retrieval evaluation is less discussed in the Cranfield-based Laboratory Research Framework for IR. Realistically one may expect between 20-40 assessments done (Borlund, 2000), out of which some are highly, fairly or marginally relevant (Sormunen, 2002). The naturalistic judgment might thus result in quite few relevant documents that are available for evaluation purposes per search job. However, as Sanderson and Zobel pointed out above one may go for a shallow layer of documents to be measured, if there are enough of search jobs performed (Sanderson and Zobel, 2005), i.e., greater than 60 jobs from a statistical point of view.

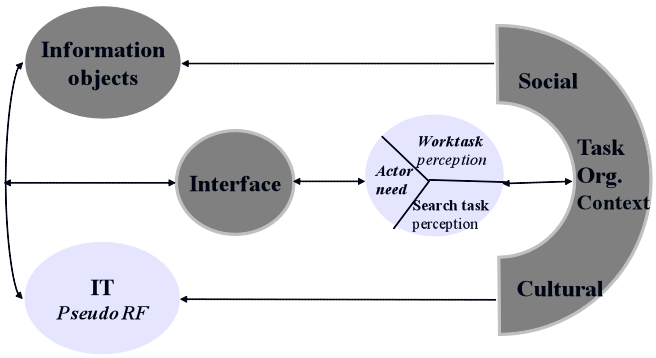

Fig. 4.7: Focus elements (in *italics*) of IR Interaction Ultra-light Laboratory study made by Kelly and Fu (2007); Kelly et al (2005).

4.3.2.1 Example Illustrating IR Interaction Ultra-Light Studies

We have chosen the Kelly and Fu (2007) and Kelly, Dollu, and Fu (2005) laboratory study of query expansion made from data extracted from the searcher's situational context. Kelly et al.'s hypothesis is outlined above, Section 2.1, as an example of such; the central IR components are displayed in Fig. 4.7.

Kelly et al.'s hypothesis was that if evidence from his/her knowledge state or/and work task description could be extracted from the searcher's task situation that evidence may improve retrieval performance, compared to the application of the request formulation only, *and* compared to various kinds of pseudo RF. Fig. 4.7 demonstrates the three variables in focus (in *italics*): 1) The IT component with pseudo RF algorithms and the searching actor component, in particular the variables of the 2) Work task perception and 3) Information need situation.

The research setting consisted of 13 test persons supplying 45 natural topics to HARD TREC, that is, topic title and description. The same persons also made the relevance assessments for their own topics as TREC assessors. The HARD TREC collection (Buckland and Voorhees, 2005) and the Lemur system with the BM25 probabilistic retrieval model was used to run a bag-of-words retrieval run for each topic based on topic title and description terms. That served as the baseline of the study. Then a number of pseudo RF models were run on top of the baseline (using top-5; top-10 ... documents for pseudo RF).

In addition the 13 test persons were asked 4 questions via an online form:

- Q1: state the times in the past you have searched that topic;
- Q2: describe what you *already know* about the topic (knowledge state);
- Q3: state *why* you want to know about the topic; and
- Q4: add any *keywords* that further describe the topic.

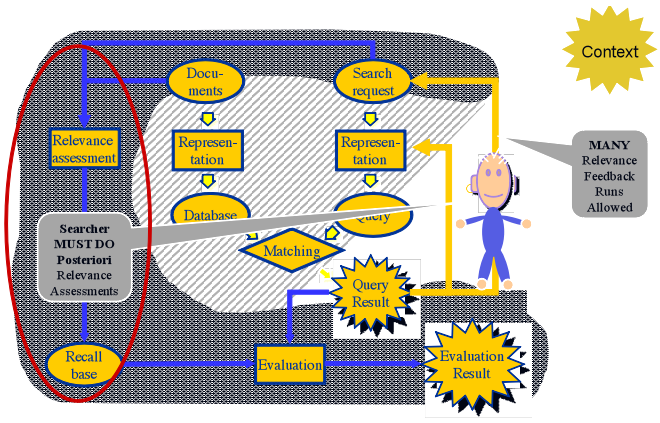

Fig. 4.8: IR Interaction 'light' – revision of (Ingwersen and Järvelin, 2005, p. 5).

The BM25, the HARD collection and the 45 topics served as controlled and neutralized variables. *Pseudo RF* variations as well as Q2-Q4 terms – on top of the baseline – served as independent variables. MAP with statistical significance test was used as dependent variable.

The results are statistically significant (t-test) and very promising from retrieval performance as well as cognitive framework points of view. The different Q2-Q4 and the baseline request yield quite different volumes of keys:

1. Baseline request: 9,33 keys
2. Q2, knowledge state: 16,18 keys
3. Q3, work task: 10,67 keys
4. Q4, added keys: 3,3 keys

The keys repeated in the various Q-forms were weighted when they were combined. Single query forms, based on the individual Q-versions, outperformed the request-based baseline. Pseudo-RF outperformed the baseline and single Q-forms. However, Q2+Q3 (and Q2-Q4 combined) outperformed any pseudo-RF on top of the baseline. This result implies that by involving the searcher situational *context* one may indeed improve retrieval performance in best match environments. The study also showed that performance increases with query length.

4.3.3 IR Interaction Light

When the IR interaction Ultra-light retrieval scenario is extended into more than one run, in which the test persons may observe the documents (or representations), it turns into an IR interaction Light laboratory study or field experiment.

In IR interaction light investigations (Fig. 4.8), the test persons themselves must carry out the relevance assessments. The scenario thus has a similar disadvantage as the ultra-light research design, in that the test persons may become saturated and produce a limited number of assessed (and relevant) documents. But the 'light' scenario is more realistic in terms of runs over a retrieval session and other behavioral patterns (including the saturation issue).

Again there are two basic research design scenarios. One is to execute a laboratory study, taking into the laboratory the test persons, as in the ultra-light studies. The assigned search jobs could be all the variations discussed above, but the original test collection assessments cannot be used. *New relevance assessments* must be applied to the search situations in a posteriori manner.

The second scenario moves the setting out into the field, e.g. into an organization, but introduces some experimental component, for instance, a novel search engine or interface configuration. This scenario is named Field Experiment in the Integrated Cognitive IR Research Framework. Obviously, the documents (database) and the context is the local one; but one wishes to try out some novel feature in that natural environment. Again, the relevance assessments must be done by the test persons themselves. In common to both alternative research designs both natural as well as assigned search jobs may be used.

4.3.3.1 Example Illustrating IR Interaction Light Studies

To illustrate a laboratory study of IR interaction light the Borlund investigation concerned with testing the ability of simulated work task situations to substitute natural information needs in evaluation of interactive IR systems (Borlund, 2000) was chosen. Her original research question was: Can simulated information needs substitute real information needs? And if yes: What makes a 'good' simulated situation? The experimental setting included the historical Financial Times collection (TREC) supplemented by an *up-to-date collection* of the Glasgow news paper The Herald. The test system was based on a probabilistic retrieval model. The design included 24 university students, graduates and undergraduates from different departments and each test person provided one real need plus was assigned 4 simulated work task situations (cover stories). The research design thus included 24 natural need situations and 96 simulated ones, which were searched by the test persons in

Natural Work Tasks (WT) & Org	**Natural Search Tasks (ST)**	**Actor**	**Perceived Work Tasks**	**Perceived Search Tasks**
WT Structure	*ST Structure*	Domain Knowledge	Perceived WT Structure	*Perceived Information Need Content*
WT Strategies & Practices	*ST Strategies & Practices*	IS&R Knowledge	Perceived WT Strategies & Practices	Perceived ST Structure / Type
WT Granularity, Size & Complexity	*ST Granularity, Size & Complexity*	Experience on Work Task	Perceived WT Granularity, Size & Complexity	Perceived ST Strategies & Practices
WT Dependencies	*ST Dependencies*	Experience on Search Task	Perceived WT Dependencies	Perceived ST Specificity & Complexity
WT Requirements	*ST Requirements*	Stage in Work Task Execution	Perceived WT Requirements	Perceived ST Dependencies
WT Domain & Context	*ST Domain & Context*	Perception of Socio-Org Context	Perceived WT Domain & Context	Perceived ST Stability
		Sources of Difficulty		Perceived ST Domain & Context
		Motivation & Emotional State	*Independent Variables*	

Table 4.3: Independent variables in the Borlund IR Interaction Light laboratory study (Borlund, 2000) mapped on to the Integrated Cognitive Research Framework dimensions.

the system located in the computing laboratory of Glasgow University. An example of one of the simulated situations from the study is given in Fig. 4.3.

Among the study's independent variables were Natural Work/Search Task – i.e., the test persons' own need situation which may have many (unknown) values; and the Perceived Search Task (information need) Contents – i.e., selected variations in kind of contents of the cover stories, such as, local topical vs. historical topical contexts, Table 4.3. All the variables in the Interface, IR algorithms, and Database dimensions, Table 4.2, were controlled, whilst the entire Access & Interaction dimension might hold potential hidden variables outside the control of the researcher, since the interaction was performed by the test persons. Latin square design and permutation of search jobs were incorporated in the design. One notices that each analysis cell in the result

matrix would contain 24 entities, just on the borderline for the application of strong significance tests like the t-test.

Both pre and post-search interviews as well as transactions logs were performed, the retrieval system became demonstrated to each test person and each person executed one training search task prior to experiments. The results showed that no difference could be found between real, natural information need situations and assigned simulated work task situations as to search runs; average use of search terms; application of different search terms; full-text-based relevance assessments; and title-based assessments. The only difference was in search time between the two kinds of search jobs. In 87 % of the search events the test persons found the simulated task situations realistic. The conclusion is that simulated work task situations (cover stories) indeed *can* substitute natural information needs. It was also found that a mixture of simulated and real task situations is applicable. The study by Borlund (2000) outlines and recommends characteristics as to what signifies 'good' simulated task situations (Borlund, 2003b), for instance, that the database must be up-to-date and the assignments realistic. Test persons are less motivated by 'historical' assignments, implying that some TREC test collections hold too 'old' materials to be used directly in IR interaction light or ultra-light studies.

Increasingly IR interaction light investigations take place world-wide, either in order to test results from laboratory simulations of real life searching behaviour or to obtain novel insight into natural IR interaction processes and behaviour – carried out in a controlled and realistic manner – see examples in the classified bibliography at the end of this chapter. Essentially the researcher should isolate very few independent variables, since natural IR interaction is complex, and attempt a robust research design. If mixed with real information situations the simulated ones may be checked for realism in their execution. The test persons must always provide the relevance assessments, which are feasible according to a grading scheme.

4.3.4 Naturalistic Field Investigations of IR Interaction - Exemplified

Naturalistic field investigations of IR interaction take two forms. They are made as *field experiments* in a natural setting, e.g. in an organizational or cultural social context, testing a novel retrieval feature or they are *field studies* that investigate searcher behaviour, satisfaction with the retrieval context or overall performance of systems. Usability or performance measures are applicable, like in IR interaction light studies.

The example of naturalistic IR interaction refers to the Marianne Lykke Nielsen (now Marianne Lykke) investigation made in an international pharmaceutical company (Lykke, 2004). Her *research goal* was to observe whether

You are product manager working for Lundbeck Pharma. A physician, who wants to know if the combination of Citalopram and Lithium leads to approve therapeutic effect on bipolar disorders, has consulted you. You need to find reports or articles investigating interaction and effect of the two drugs.

Fig. 4.9: Simulated work task situation assigned the test persons by Lykke (2004).

a company thesaurus (local ontology) based on human conceptual associations affects searching behaviour, retrieval performance, and searcher satisfaction different from a domain-based thesaurus. Already a few years previously she had developed the *associative thesaurus* by means of association tests made by 35 research and marketing employees from the company (Lykke, 2001). The test persons provided synonyms, narrow and broader concepts founded in the 'company vocabulary'. The associative thesaurus was slightly larger in number of entries (379 entries more) than the *domain thesaurus*, made by domain experts and based on a 'scientific vocabulary'. The latter ontology served as the control mechanism in the later experiments.

Creating the associative thesaurus was a kind of field study, whilst the investigation of its effects on searcher behaviour and IR performance became a *field experiment*. 20 test persons from the basic and clinical researchers, including marketing staff also with pharmaceutical background, were each assigned three simulated work task situations. They all had the same task structure and level of complexity, and were based on real work tasks observed via recently logged requests to the company retrieval system, see Fig. 4.9.

Blind testing was used in the research design: The test persons were told that the investigations were part of the system design process. They did not know which thesaurus type they were actually interacting with – only the research team knew that. Latin square design to avoid learning effects was used with permuted sequences of search jobs given the subjects. The 20 test persons would consequently in total make 30 searches in each thesaurus system. Relevance assessments were made via a three-grade scale: Highly, partially, and not relevant. Recall and precision served as performance measures – aside from the use of satisfaction measures and other behavioural observations.

There were some constraints usual for research designs in commercial environments. Only two working hours per test person was allowed by the employer. This time slot covered capture of search skills (actually done via e-mail prior to the allocated period); explanation of the research setting; pre-search interview of searcher's mental model of each search job plus capture of expectations; search session of the three search jobs with relevance assessments of retrieved documents; and post-search interview of motivations and satisfaction for each search job.

Fig. 4.10 displays the independent variable, Thesauri, with two values: the associative thesaurus (ASSO) and the domain-based one (DOMAIN).

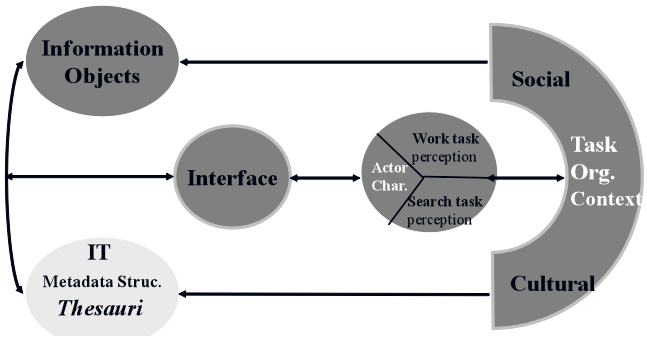

Fig. 4.10: Independent variables (*italics*) and potential influential variables (white characters)

With reference to Tables 4.1-4.2, the controlled variables were the Natural Work/Search Task dimensions (the organizational context of Lundbeck Pharma); Perceived Work Task Structure, Complexity (high); Perceived Search Task, Information Need content; Documents and Sources; Retrieval Engine; and Interface dimensions. Hidden (or influential/modifying) variables could adhere to the Access & Interactions dimension, but also to the Actor characteristics.

Among the interesting results were: both thesauri demonstrate quite the same performance in recall & precision (ASSO: .14 & .32; DOMAIN: .11 & .37). Both thesauri were applied to query formulation and modifications and did provide lead-in terms to searching. The time using ASSO was slightly longer than using DOMAIN, the latter was more used in the pre-search stage. Quite interestingly the test persons assessed the same documents quite differently. This (unexpected) phenomenon owed to the fact that the two major staff groups (basic vs. clinical/marketing researchers) used the two thesauri very differently. This difference in Actors was found to be a *hidden (influential) variable* in the study. Fortunately, the distribution of the two group members over the research execution was equal and did not interfere with the results. Basic researchers appreciated ASSO because they could easier explore new drugs and use local and novel vocabulary; clinical staff preferred DOMAIN for their rigorous clinical and standard scientific drug testing purposes.

The naturalistic IR interaction example demonstrates how complex the research design may become when the control of the context is slackened, although as many dimensions of variables as possible actually are either neutralized statistically (the simulated tasks) or directly under control. Both IR interaction light and the naturalistic investigations *per se* allow for freedom of the Access & Interaction variables; they cannot be so easily controlled in experiments directly aiming at those processes.

4.4 Conclusions

In pure laboratory experiments following the Laboratory Research Framework for IR only *simulations* of searcher behaviour can be executed. If existing test collections with sets of assigned 'topic' and corresponding relevance assessments are used in interactive IR investigations only *IR interaction ultra-light* studies are feasible. In such short term IR investigations the single encounter between the test person and retrieved documents assures the avoidance of learning effects – in line with the conditions of human test collection assessors. If non-test collections are applied in the investigations, i.e., that assigned as well as natural search jobs are feasible, one may chose between ultra-light studies or the session-based *IR interaction light* research design. In both research settings the test persons *must* perform the relevance assessments and other usability measures.

IR interaction ultra-light experiments are tightly controlled, but less realistic owing to the short term interaction. The interaction light studies are less controlled but more realistic, although both takes place in the laboratory. The ultra-light laboratory experiments in particular are highly effective in tightly controlled IIR investigations, as demonstrated by Kelly and Fu (2007) and Kelly et al (2005). Ultra-light IR interaction investigations are thus recommendable for computer scientists who wish to try out a first step of IR interaction. Meanwhile, the number of test persons, search jobs and set-ups are the same for the two research designs, as well as for the third step into context: the IIR *field experiments* and studies. The advantages of interaction (ultra)-light and field experiments are the freedom of choice between assigned requests, assigned simulated work task situations and/or natural or real tasks. The search jobs may be associated with particular working or organizational contexts or with daily-life situations. Another advantage is of methodological nature. In all IR interaction studies several *data collection* methods are applicable simultaneously. Recall base evaluation, client logs, observation (including eye-tracking) and interviewing provide in combination valuable information on the interaction phenomena and performance.

Another central feature of IR interactive investigations concerns the relevance scaling, which have moved beyond the binary one into more realistic 3-4 graded relevance scales (Kekäläinen, 2005; Kekäläinen and Järvelin, 2002b; Sormunen, 2002), at the same time as novel performance measures have been defined and tested, such as the DCG family (Järvelin and Kekäläinen, 2002; Ingwersen and Järvelin, 2005).

The *disadvantage* of the IR interaction (ultra)-light studies and field experiments lies in the relatively small number of assessed documents in realistic settings in which the test persons are allowed to perform 'natural', on their own terms and are becoming saturated. The balance between having control of the research setting and, at the same time, allowing for a certain degree of realism is vital for the validity of interactive investigations. Thus, when realism dictates that the research designs operate with relative few relevant

documents and that performance hence should be measured at a shallow depth of the ranked result lists, novel measures are called for. One way would be the suggestion made by Sanderson and Zobel (2005) among others to allow for shallow-level measures but then to measure over more IR interaction events.

When stepping into context outside the laboratory the number and complexity of variables increase. The Laboratory Research Framework for IR, Fig. 4.1, does not display many variables. With the exception of the assessor they are all rather tightly controlled. The Integrated Cognitive Research Framework for IR offers nine dimensions of variables including those of the former framework. Although the number and complexity of variables has increased in the Integrated Cognitive Research Framework it also offers methodological tools for handling the higher level of complexity and suggestions to explore a long range of IR phenomena involving searchers. In particular, the strength of the Integrated Cognitive Research Framework lies in its capacity of pointing to potentially hidden or influential variables in investigations, and how to neutralize them in research designs.

Classified Bibliography

IR Interaction 'Light'

- Beaulieu (1997)
- Beaulieu (2000)
- Beaulieu and Jones (1998)
- Belkin et al (1993)
- Belkin et al (1996a)
- Belkin et al (1996b)
- Belkin et al (2003)
- Bilal (2000)
- Borlund (2000)
- Byström and Järvelin (1995)
- Borlund and Ingwersen (1998)
- Campbell (2000)
- Cothey (2002)
- Efthimiadis (1993)
- Joachims et al (2005a)
- Koenemann and Belkin (1996)
- Maglaughlin and Sonnenwald (2002)
- Palmquist and Kim (2000.)
- Petrelli et al (2004)
- Puolamäki et al (2005)
- Ruthven et al (2002)

- Ruthven et al (2003)
- Tombros et al (2003)
- White et al (2003)

IR Interaction 'Ultra-Light'

- Belkin et al (1982)
- Harper et al (2004)
- Hsieh-Yee (1998)
- Kelly and Fu (2007)
- Papaeconomou et al (2008)
- Kelly et al (2005)
- White et al (2005a)
- White et al (2006)

Laboratory Experiments IR simulations

- Croft and Thompson (1987)
- Dennis et al (2002)
- Magennis and van Rijsbergen (1997)
- White (2006)
- White et al (2005b)

Methdological Issues Research Design

- Baeza-Yates and Ribeiro-Neto (1999)
- Belew (2000)
- Efron (2009)
- Ericsson and Simon (1996)
- Fidel (1993)
- Fidel and Soergel (1983)
- Frankfort-Nachmias and Nachmias (2000)
- Hawking et al (2000)
- Hersh et al (1996)
- Ingwersen and Järvelin (2005)
- Järvelin (2007)
- Ingwersen and Järvelin (2007)
- Borlund (2003b)
- Belkin et al (1983)
- Harman (1996)
- Ingwersen and Willett (1995)
- Kekäläinen and Järvelin (2002a)

- Saracevic et al (1988)
- Tague and Schultz (1988)
- Teevan et al (2004)

Naturalistic IR Interaction

- Anick (2003)
- Barry (1994)
- Bellotti et al (2003)
- Bilal (2002)
- Dumais et al (2003)
- Fidel et al (1999)
- Ford et al (2001)
- Hirsh (1999)
- Lykke (2004)
- Lykke (2001)
- Kelly and Belkin (2004)
- Saracevic and Kantor (1988a)
- Saracevic and Kantor (1988b)
- Vakkari (2001)
- Vakkari and Hakala (2000)
- Wang et al (2000)
- Wang (1997)

Relevance and Performance Measures Issues in Interactive IR

- Barry and Schamber (1998)
- Borlund (2003a)
- Cosijn and Ingwersen (2000)
- Czerwinski et al (2001)
- Frokjaer et al (2000)
- Hornbaek (2006)
- Kekäläinen (2005)
- Kekäläinen and Järvelin (2002b)
- Käki (2004)
- Järvelin and Kekäläinen (2002)
- Järvelin and Kekäläinen (2000)
- Larsen and Tombros (2006)
- Lun (2001)
- Nielsen (2003)
- Salojärvi et al (2003)
- Sanderson and Zobel (2005)
- Spink et al (1998)

- Vakkari and Sormunen (2004)
- Voorhees (1998)