ICTIR, Budapest 2007,
Proceedings, p.135-147.

ON THE HOLISTIC COGNITIVE THEORY FOR INFORMATION RETRIEVAL

Drifting Outside the Cave of the Laboratory Framework

Peter Ingwersen
Royal School of Library and Information Science, Birketinget 6, DK 2300 Copenhagen S, Denmark
pi@db.dk.

Kalervo Järvelin
Department of Information Studies, University of Tampere, Finland
Kalervo.jarvelin@uta.fi

Keywords: Information retrieval theory; Cognitive Research framework; Laboratory Framework; IR interaction; Relevance in IR; Research variables.

Abstract: The paper demonstrates how the Laboratory Research Framework fits into the holistic Cognitive Framework for IR. It first discusses the Laboratory Framework with emphasis on its underlying assumptions and known limitations. This is followed by a view of interaction and relevance phenomena associated with IR evaluation and central to the understanding of IR. The ensuing section outlines how interactive IR is viewed from a Cognitive Framework, and 'light' interactive IR experiments are suggested performed by drawing on the latter framework's contextual possibilities. These include independent variables drawn from a collection, matching principles in a retrieval system, and the searcher's situation and task context. The paper ends with concluding points of summarization of issues encountered.

1 INTRODUCTION

According to Järvelin (2007), since the year 2000 the Call for Papers for the ACM SIGIR Conference has not mentioned *information retrieval (IR) theory* as one of the key areas for which submissions are called. In 2000, papers were called for "IR Theory, including logical, statistical and interactive IR models, and data fusion", among others. Since then, the corresponding item in the Call for Papers has been, more often than not, "Formal Models, Language Models, Fusion/Combination". This also suggests how *theory* is to be interpreted in the SIGIR context. Studies on interaction or users belong under another heading, typically consisting of the subheadings "Interactive IR, user interfaces, visualization, user studies, and user models" (p. 970).

In the present paper we discuss selected aspects of an IR theory, that is, the holistic Cognitive Framework for IR as put forward and analyzed by Ingwersen and Järvelin (2005) in association with the Laboratory IR Framework. The selected aspects are *interaction* and *relevance* since they are central to understanding IR and form the vortex of any scientific perspective of IR research.

By theory we understand systematic collections of theoretical and empirical laws and associated existence assumptions. A theory explains observed regularities and hypothesizes new ones. Further, a theory provides deeper understanding of phenomena by using theoretical concepts that go beyond immediate observations. Hence, scientific theories represent reality, systematize knowledge concerning it, and guide research (e.g., by suggesting novel hypotheses) (Järvelin, 2007; Bunge, 1967).

The holistic cognitive theory for IR originated 1990-1992 and became more profoundly analyzed in (Ingwersen, 1996) leading to an increasing weight of empirical studies based on hypotheses derived from that theory (Ingwersen and Järvelin, 2005). It replaced a more individualistic perspective of the cognitive view in (interactive) IR dominant from

mid-1970. Among the hypotheses is the principle of polyrepresentation (Larsen, et al., 2006; Skov et al., 2006; White, 2006), alternative ways of understanding information and relevance (Borlund, 2003), and a novel research framework (Ingwersen and Järvelin, 2005, p. 313-358). In this presentation we focus on elements of this framework that incorporates the perspectives of cognitive IR theory *and* the Laboratory Framework for IR. It attempts to provide more potential for hypothesis generation, research design and execution in IR than the Laboratory Framework by adding *context* to it. Eventually, the framework may lead to a research program for IR (p. 359-376).

In line with D.C. Engelbart (1962, p. 2) we see the Laboratory Framework, Figures 1 and 3, as well as the cognitive models for IR, Figures 4-5, as *conceptual frameworks* that specify:

- Essential objects or components of the system to be studied.
- The relationships of the objects that are recognized.
- The changes in the objects or their relationships that affect the functioning of the system.
- Promising or fruitful goals and methods of research.

With Järvelin (2007) we define the concept *model* to refer to a precise (often formal) representation of objects and relationships (or processes) within a framework, as in the probabilistic IR Model. In principle, modeling also may involve modeling human actors and organizations, e.g., as done in the relevance assessment models, Figures 2 and 6.

Already in 1992 Robertson and Hancock-Beaulieu started the discussion of the so-called three revolutions in IR: the interactive, the relevance and cognitive ones, the latter being a consequence of the two former revolutions. IR became regarded a process that leads to human perception of information, interpretation, learning and cognition. This coincided with the view of (interactive) information retrieval and seeking research as divided into the traditional mainstream system-oriented laboratory-based line of IR research, i.e., the algorithmic perspective of IR or the Cranfield paradigm, *and* the realistic user-centered interactive information seeking and retrieval approach (Ingwersen, 1992). Even though this binary division made the research situation quite clear-cut it did not provide any support from the one approach to the other and did not improve the research environment.

The situation was not solved at all mainly because no mutual foundation between the two tracks of research could be established. The holistic cognitive framework attempted to do exactly that from Ingwersen (1996). The reason why it may succeed, and probably already has contributed positively to a more relaxed attitude among IR researchers to new forms of design and evaluation settings, is that all parties somehow agree that the ultimate goal of information retrieval is to facilitate human access to and *interaction* with information, in whatever form, that probably may entail cognition. In return, the better the seeking actor can be supported to support the IR system, the better the overall retrieval performance. These ideas of mutual support across research perspectives can be seen to build the foundation for the increasing number of empirical studies and experiments on implicit and explicit relevance feedback (RF), searcher's task descriptions, use of simulations, recommender systems and personalization of retrieval.

In the monograph by Ingwersen and Järvelin (2005) the differences between the Laboratory Framework and the holistic Cognitive approach are listed and discussed (p. 192-194). They range from conception of information; task dependency; IR system setting; role of intermediary; over context; conceptual relationships; into evaluation approaches. Two of the central differences concern *interaction* and *relevance* conceptions, and hence, are concerned with research design including IR evaluation. We intend to demonstrate that the two research frameworks fit together and in symphony contribute to an improved understanding of both phenomena.

The remaining paper first discusses the Laboratory Framework with emphasis on its underlying assumptions and known limitations. This is followed by a discussion of interaction and relevance phenomena associated with IR evaluation and central to the understanding of IR in an extension of the Laboratory Framework towards context. The ensuing section outlines how interactive IR is viewed from a Cognitive Framework, and 'light' interactive IR experiments are suggested performed by relying on the latter framework's contextual possibilities. These include independent variables drawn from a collection, matching principles in a retrieval system, and the searcher's situation and task context. The paper ends with concluding points of summarization for of issues encountered.

2 THE LABORATORY IR FRAMEWORK

The Laboratory Framework is shown in Figure 1, the so-called cave perspective owing to its almost context-free nature. According to Järvelin and Ingwersen (2005, p. 4-6) in this perspective an IR system consists of a database, algorithms, requests, and relevance assessments made by assessors, stored in the recall base. The system components are represented in the middle and the evaluation components on top, left and bottom in the shaded area. The main focus of the research has been on document and request representation and the matching methods of these representations.

In this view real users and tasks are not seen as necessary. They are not needed for testing the matching algorithms for the limited task the algorithms are intended for: retrieval and ranking of topically relevant documents. Test requests typically are well-defined topical requests with verbose descriptions that give the algorithms much more data to work with for query construction than typical real life IR situations (e.g., web searching) do. However, recently the TREC (Text REtrieval Conference) experimental environment has been extended to involve a Web track with realistically short requests.

The rationale of evaluating the algorithmic components consists of the goals, scope and justifications of the evaluation approach. With reference to Ingwersen and Järvelin (2005, p. 6-7) the *goal* of IR research is to develop algorithms to identify and rank a number of topically relevant documents for presentation, given a topical request.

Research is based on constructing novel algorithms and on comparing their performance with each other, seeking ways of improving them, between competitive events. On the theoretical side, the goals include the analysis of basic problems of IR (e.g., the vocabulary problem, document and query representation and matching) and the development of theories and methods for attacking them.

The *scope* of experiments is characterized in terms of types of experiments, types of test collections and requests as well as performance measures. The experiments mainly are batch-mode experiments. Each algorithm is evaluated by running a set of test queries, measuring its performance for individual queries and averaging over the query set. Some recent efforts seek to focus on interactive retrieval with a human subject, the TREC interactive track being predominant, Figure 3. The major modern test collections are news document collections, only recently complemented by Web repositories and some collections of objects from other media. The major performance measures are recall and precision.

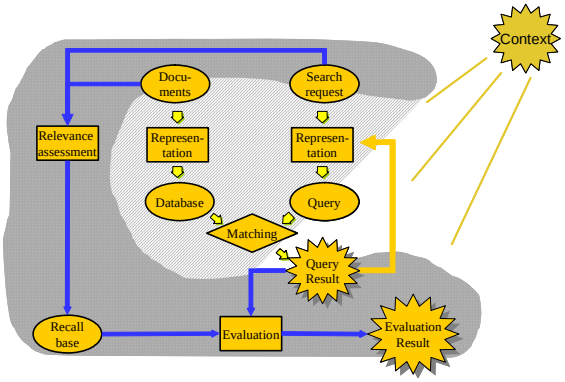

Figure 1: The Laboratory Research Framework schematized (Revision of Ingwersen and Järvelin 2005, p. 5).

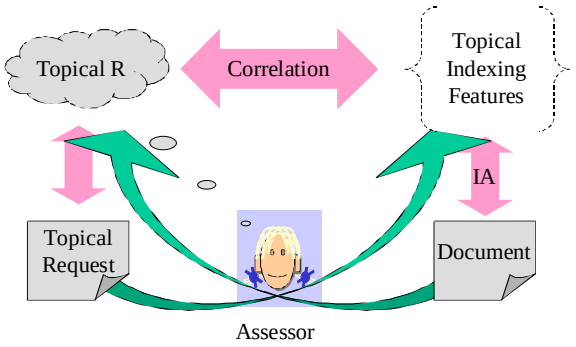

Figure 2: Justification of the Laboratory Framework (Kekäläinen and Järvelin, 2002a).

With reference to the Framework, Figure 1, we observe that the roof of the cave, document and request types/genres, could be more extensively explored *within* the framework boundary. Similarly, the relevance assessment process is not undergoing extensive study, except for a few investigations of neutralizing by statistical means the deviations of assessments in the case of introducing several assessors in a test collection (Vorhees, 1998) or in terms of re-assessments into graded relevance (Sormunen, 2002). The latter investigation can also be seen as check-up of the original binary assessment quality made by the collection assessor in TREC7-8. Further, the recall base necessitates that no other human actor (aside from the assessor) participates, in order not to make the assessments unusable.

However, experimental design allows for several retrieval iterations (runs) with feedback from the system and automatic 'pseudo RF' along the vertical arrow at the opening of the cave – simulating a primitive searcher.

2.1 Justification, Goal and Scope of the Framework

The *justifications* of the Framework may be discussed in terms of Figure 2. The main strength is that words and other character strings from texts, when distilled as indexing features by an indexing algorithm (IA), correlate, with fair probability, to the topical content of the documents they represent and to the queries which they match (save for problems of homography). When a test user (or algorithm) processes a topical request, it is possible to predict, with fair probability, which indexing features should be considered (save for problems of synonymy, paraphrases). Because the topical request also suggests topical relevance criteria, there is a fair correlation (clearly better than random) between the indexing features of matching documents and a positive relevance judgment. Indexing features correlate to meaning in the topical sense. The more features that can be used as evidence, the better the retrieval.

This observation is crucial when understanding the success of the Laboratory IR Framework in text retrieval. All indexing features are regarded independent from one another in most logic-based and statistical retrieval algorithms. When dependence has been considered, it has not been found to significantly improve retrieval results – suggesting that the key to success lies elsewhere. Although intuitively unrealistic (after all, authors commonly put together words intentionally and meaningfully) text retrieval succeeds exactly because many content bearing features (words) brought together has a greater chance of hitting some (few) meaningful portions of information objects (texts) than only few features do. Few features hit too many objects. Although text writing is a non-chaotic but stochastic process (Egghe and Rousseau, 1990) as to 'function words' (Bookstein and Swanson, 1974), this is not the case for 'specialty words', that is, content bearing words that are informative or discriminating about the document contents. Such words are not randomly distributed but follow a pattern organized by the thematic progression of the text (Katz, 1996, p.16). Hence, there exists a cognitively associated explanation with respect to why the statistical algorithms function well.

2.2 Limitations and Extension of the Framework

First of all the Laboratory Framework is almost context-free. The sole contextual matters are that documents and requests are supposed to derive from some sources or actors external to the cave, metaphorically speaking – yet see the Cognitive Framework, Figure 5. In addition, the assessments are made by some assessors supposedly mirroring (as a kind of shadow on the cave wall) real searchers. The matching algorithm is physically located in the cave and thus put there by some designer/researcher – during experiments external to the cave but acting as an implicit context and observer.

Figure 3 displays the Framework with kinds of explicit and implicit relevance feedback possibility offered by a human participant acting as information searcher at the right-hand side. This shape of the Framework including the searching actor is mandatory if probabilistic-like IR Models are to work properly, owing to the necessity of RF for calculating proper probabilities of relevance for the documents in the collection. The Framework then allows for two retrieval runs as the maximum; exceeding that number the recall base becomes unusable: one initial run made automatically from a request set; and one consecutive run made during the same session based on RF made by the searcher (the outer vertical arrow). The searcher is simply instrumental to providing relevance odds and is supposed to understand the request and query result the same way as intended by the assessor used by the test collection designer – see also below for modes of performance measurements.

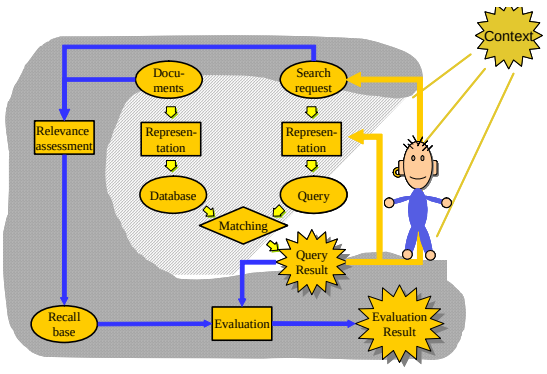

Figure 3: The extended Laboratory Framework for IR.

By this extension of the original Laboratory Framework performance is then measured after the second retrieval run against the first run, now seen as a baseline. This allows for a liberal variety of experiments with RF and weighting methods in query modification algorithms. We have now four ways of measuring performance of one session: a) by an assessor in a pre-existing text collection (via pooling of results from the first run across all competing IR engines, as in TREC); b) by a searcher-independent assessor judging the second run result across all competing engines; this latter mode is like using pseudo RF, comparing to mode a) scores; c) by one test searcher of the second retrieval run result (the first run' relevance result also exists in logs); d) by all searchers of the second run of the same query session. In all cases one may pool the performance scores across the set of request/queries given in the experiment.

However, more than one consecutive retrieval runs made or assessed by the searcher are not allowed, since learning effects may alter the searcher's perception of both new (and already seen) documents and thus change their relevance away from the fixed ones in the recall base. Hence, we observe a severe limitation of the Framework seen from a realistic point of view. On the other hand one may state that performance measurement modes c) and d) are truly user-based. There may, however, appear variability between second run relevance scores owing to different searcher perceptions of the same first run result, mode d), implying a *cognitive drift* and hence uncertainty as to the comparison to the recall base relevance scores. Therefore, pooling across the test searchers and sessions may help *neutralizing the effect* of such individual perceptions. But they do not disappear though! This also provides an opportunity for the experimental setting to apply graded relevance by means of averaging the scores per document. This presupposes that 1) all test persons are forced to assess the same (number of) documents; or that 2) they assess a substantial portion so that enough scores are provided per document to make the averaging reasonable. The realistic liberty of choice is hence somewhat limited.

Measurement modes a) and b) provide more consistent assessments, in particular if the original requests (= TREC topics) are generated by the same assessor judging the retrieved documents. This might not be the case in the modes c) and d) since the requests may be given to the test searchers. Although we are still within the Laboratory Framework the conditions for the experimental situation now resembles the more open Cognitive Framework for IR, Figure 4: the searchers may either be given a simulated work task situation (or cover story) to start the session from, interpreting their own (realistic) information situation, relevance and need from the simulation; or they provide their own information situation to be solved Borlund (2003b). In both cases the recall base (and the classic test collection conception) stops fitting the experimental setting. The performance measurements are then done as in modes c) or d) across all the runs per session and sessions – the former mode c) being less controllable due to the retrieval freedom allowed in field experiments. One may, just for fun, compare what the test searchers achieve of performance scores with the assessor scores for the same requests. The likelihood is that very rarely will the two scores be similar, albeit that they should never be compared in the first place for logical reasons. Even a direct and allowed comparison of one-run results over several different assessors provides distinctive different results (Sormunen, 2002). Obviously, modes c) or d) have to be applied if the experiment addresses interface issues or document presentations and not simply document ranking principles.

2.3 Relevance and Interaction in the Extended Laboratory Framework

We can make two central observations concerning the extended Laboratory Framework.

First, *relevance* is taken as topical, but factual features (based on data items, like author names and other metadata features) could be included. Relevance also is static between a topical request and a document as seen by an assessor. This observation is necessary according to the discussion above of the evaluation measurement modes. Further, like for document features the assessments are independent of each other (i.e., no learning effects, no inferences across documents may or can occur) and there are no saturation effects (i.e., in principle the laboratory assessors do not get tired of repetition). The assessors do not know, in which order the documents would be retrieved (owing to the pooling) so they cannot do otherwise or properly model user saturation. Relevance is commonly also binary. But recent developments, also under influence of empirical results (Borlund and Ingwersen, 1998, Borlund, 2000; 2003b) adhering to the Cognitive IR Framework, Figure 4, have made it possible to include graded relevance in experimental settings within the limits of the Laboratory Framework and

generalize the performance measures (Järvelin and Kekäläinen, 2000, 2002; Kekäläinen and Järvelin, 2002b;). Similarly, experiments on highly relevant information objects have been introduced (Vorhees, 2001; Kekäläinen, 2005) – lately leading to the HARD track in TREC.

Secondly, *IR Interaction* presents a similar case as for relevance. Realistic complex dynamic interaction is regarded but a sequence of simple independent topical interactions. The problems encountered above with searcher RF during interactive IR do not seem to influence this basic assumption within the Laboratory Framework. Henceforward, good one-shot performance by an algorithm should be rewarded in evaluation according to the Framework. Changes in the user's understanding of his information situation and need should thus affect the consequent request and query, but in the Framework seen as a completely *new* retrieval situation. This view is somehow in contrast to the mandatory application of RF over two connected retrieval runs, in order to reduce uncertainty by statistical means in probabilistic IR. One way of looking at this contradiction is that the Vector Space Model appeared first and established the Laboratory. Framework based on the Cranfield design. Vector Space does not really require human RF in order to function. One may in fact argue that the probabilistic Model (Robertson, 1977) is far more user-driven and cognitive than most other IR algorithms, in line with the cluster hypothesis based on searcher input (Willett, 1988).

A way to deal with the extended Framework but avoiding the cognitive drift and pitfalls mentioned on performance measurements c) and d) above, but in connection to IR interaction, is to *simulate the searcher* behavior during retrieval. This has been done rarely but successfully by Magennis and van Rijsbergen (1997) on query expansion, and lately by, for instance, White et al. on implicit RF with user validation (2004; 2005), Wang et al. (2006) on log-based collaborative filtering and by Keskustalo and al. on graded RF (2006). The searcher, Figure 3, is simulated by means of a user model. The most crucial parameter in that kind of research is to obtain a realistic and reliable user model, based on sound assumptions, for which all the variables are accounted for. Obviously, the easiest way to get a user model is to 1) obtain information on searchers from the empirical research done outside the Laboratory Framework, to the right of the opening of the cave, drifting into context; 2) thinking out some variables to investigate, e.g., the number of documents a person would like to look at for RF or the retrieval mode: all there is of relevant stuff vs. only highly relevant documents (or parts of XML objects). Then all possible combinations are tested and the best performing alternatives are selected to be tested in real life. This is achievable through user simulations, since any learning effects may be avoided in repeated experiments. These perspectives and solutions within the Laboratory Framework lead us logically to the Cognitive Framework for IR.

3 THE COGNITIVE IR FRAMEWORK

As discussed above, the extended Laboratory Framework evidently drifts into an increasingly cognitive and physical sphere outside the cave: the contexts located on the right-hand side, Figure 4.

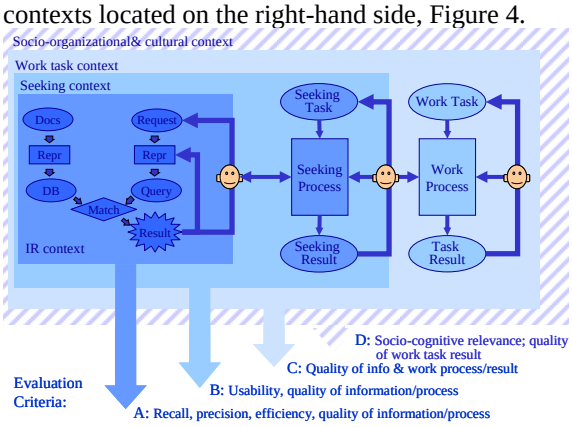

Figure 4: Nested contexts and evaluation criteria for task-based information access (extension of Kekäläinen and Järvelin, 2002a)

Algorithmic IR is here seen in context of information seeking and work task processes, job-related or not (Ingwersen and Järvelin, 2005, p. 322). Interactive processes take place horizontally whilst evaluations and RF is vertical at each level of processing. For each nested context is given the kinds of evaluation criteria that might apply to that experimental situation. The model derives from the holistic Cognitive framework, Figure 5. The central presuppositions of the Cognitive Framework are that (p. 25):

1. Information processing takes place in senders *and* recipients of messages;
2. Processing takes place *at different levels*;
3. During communication of information any actor is

influenced by its past and present experiences (*time*) and its social, organizational and cultural environment;
4. Individual actors *influence* the environment or domain;
5. Information is *situational* and *contextual.*

First, it is equally valid to the framework whether the processing device acts as a sender or recipient of signs, signals or data, for example, during communication processes. This implies that the framework not only treats human actors as recipients *but also* as generators of signs to and from machines and knowledge resources – arrows 6-8, Figure 5. Thus, there are *constantly* several human actors involved in IR as variables – not simply an algorithmic designer, a test collection assessor or a searcher.

The holistic cognitive view is consequently not limited to user-centered approaches to information. Essentially, it is *human*-oriented but encompasses all information processing devices generated by man as well as information processes intended by man. The former refers, for instance, to computers or other forms of technology; the latter signifies acts of generation, transfer, and perception of information, for instance, by technological means.

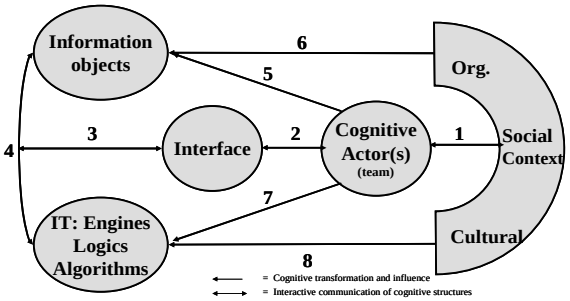

Figure 5: The holistic Cognitive Framework for IR (Ingwersen and Järvelin, 2005, p. 261).

The left-hand side of the Framework incorporates the Laboratory Framework, as in Figure 4, with arrow (2) signifying human request making and RF, arrow (3) query generation or modification and (4) the matching of documents by means of IR processes. Arrow 1) refers to social interaction between, for instance, a searcher and his/her immediate social context, i.e., information seeking without IR.

In relation to both figures, IR belongs to the searcher's information seeking context where it is but one means of gaining access to required information. This context provides a variety of information sources/systems and communication tools, all with different properties that may be used based on the seeker's discretion and in a concerted way. The design and evaluation of these sources/systems and tools needs to take their *joint usability*, quality of information and process into account. One may ask: what is the contribution of an IR system at the end of a seeking process – over time, over seeking tasks, and over seekers? (Ingwersen and Järvelin, 2005). Since the knowledge sources, systems and tools are *not* used in isolation they should not be designed nor evaluated in isolation. They affect each other's utility in context. This is the reason for the statement that all the five components of the Cognitive Framework, Figure 5, *and* the interaction process itself, are contextual to one another. One cannot see the searcher isolated from the social context, but neither is it possible to view that actor's activities without the *systemic context*: the Laboratory Framework for IR, so to speak. They influence each other. Whenever one changes, the others need to adapt to avoid dissonance – causing a dynamic imbalance in the process.

An obvious counterargument, according to Ingwersen and Järvelin (2005, p. 323) is that there are too many seeking contexts with too many possible combinations of systems and tools: The design and evaluation of IR systems becomes unmanageable. Therefore it is best to stick to the Laboratory tradition of design and evaluation. If one does not know more than one's own unsystematic recollection of personal IR system use, such design and evaluation demands may be of tall order, indeed. However, even limited knowledge on real information access may reveal typical uses, strengths and weaknesses of various tools and systems – and how their users perceive them. This provides a better basis for design (and simulations) than de-contextualized standard assumptions and measures. If automobile designers would behave alike, they would focus on the engines (e.g. horsepower, acceleration) no matter whether they design a sports car, pick-up or a truck!

The cognitive drift observed above associated with the Laboratory Framework, Figure 3, forces IR research to look into at least some of the variables close by the 'cave opening'. The simulated searcher behavioral research is but one example of this necessity. Real-life tests move the scenarios out in the open outside the cave, so to speak. Searcher and task contexts are becoming interesting objects for IR research, like the observation and implementation of implicit RF, quests which necessitate the involvement

of human test persons[1].

3.1 Interaction and Relevance Issues

Retrieval strategies during IR interaction rarely are addressed in the laboratory or in field experiments with best match systems or on the Web. This implies that building block strategies with facets or planned try and error-like tactics, known from online retrieval, have hardly been studied in a real-life best match environment. Owing to the primitive retrieval mechanisms in Web search engines, including Google and Yahoo, the searches are not connected. They correspond to the independency assumptions behind the Laboratory Framework for IR.

First recently structured queries are being tested during retrieval (Kekäläinen and Järvelin, 1998) with improved performance as a result, compared to common bag-of-words IR. Further, with test searchers given a topic (or a simulated situation) as starting point for retrieval it often happens that the initial query is really bad. RF may then not solve such problems, probably also because the searcher's knowledge state is weak concerning the search objective. Moreover, real searchers have different interpretations, and consequently, construct differently behaving queries – even when facing the same scenario. This does not become apparent when (verbose) topics of test collections are (automatically) used in experimentation. In real life there seldom are lengthy topical need descriptions – there rather is a multitude of possible interpretations that need to be mapped to a collection.

In the laboratory, one assumes that searches are well formulated, and searchers knowledgeable to provide RF and never getting tired of looking at retrieved objects – mirrored as an assessor. In field studies the ensuing IR interaction may result in intermediate retrieved sets of viewed documents but researchers rarely look into those sets, commonly only into the last retrieved ranking and its scores. However, intermediate feedback from the system might provide clues to a better understanding of how human query construction occurs – even when based on a simulated situation and over many test searchers.

With respect to *relevance* associated with the Cognitive IR Framework, Figures 4-5, both Saracevic (1996), Cosijn and Ingwersen (2000) and Borlund (2003a) have promoted the typology of relevance that includes lower order relevance: algorithmic and topicality; as well as higher order relevance: pertinence, situational relevance and socio-cognitive relevance. Pertinence signifies the relation between the internal information need situation and the retrieved objects. We may here talk about parameters like currency, novelty, authority, that is, features that are concerned with objects' *isness* (Ingwersen and Järvelin, 2005, p. 271) like names, dates, layout etc. These are not so difficult to deal with in IR settings. Even of higher order the socio-cognitive relevance (Cosijn and Ingwersen, 2000) is quite objective and signifies the temporal evaluations of information objects by actors, such as represented by scientific citations, quotations, inlinks or simply mentioning in objects.

The only problematic relevance type, but also the most discussed, appreciated by seeking research and theoreticians, is *situational relevance*. It is supposed to be the relationship between the work task as perceived by the searching actor and the retrieved information objects. In a cognitive sense situational relevance is problematic by being a theoretical construct that serves well in discussions of relevance issues, but when investigated constantly is illusive, not present in logs or searcher statements/interviews (Vakkari, 2001; Berry and Schamber, 1998). It may perhaps associate to indexing keys directly aiming at work tasks *added* to the objects; or it relates to features of the person's own knowledge state which are difficult to compute – see Figure 6. The concept seems in-operational. In a metaphoric way, situational relevance serves as the concept of 'dark matter' in space science.

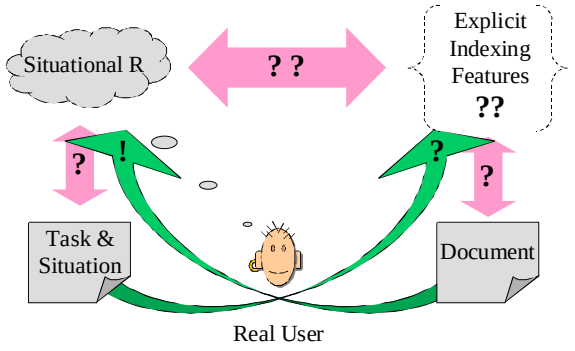

Figure 6: Situational relevance in IR (Kekäläinen and Järvelin, 2002a).

As discussed by Kekäläinen and Järvelin, 2002a) a real user, being thrown into a situation, may well be able to recognize a relevant document once presented (therefore the exclamation mark, Figure 6). However,

[1] Note the increase of context associated with e.g. the ACM-SIGIR Conferences from 2004: IR in Context workshops; or Contextual IR workshops in the Context Conference.

he/she may have difficulty in discussing the relevance criteria of the task and situation (Kekäläinen and Järvelin, 2002a). Further, he/she certainly has difficulty in expressing a request and formulating a query to the IR system, at least anything other than topical as long as text is concerned (but save for bibliographic fields etc., metadata, if available), because current systems do not provide for anything else (thus the question marks, Figure 6). The system designer probably never had any idea of other than explicit topical indexing features, because there is *no known pattern* of situational indexing features that are explicit in text – the computer does not handle implicit features – and useful to searchers. Therefore the available indexing features may not correlate to the situational relevance criteria, which the user did not express, save for one thing: topical relevance heavily correlates to situational as already suggested by Burgin (1992) and Vakkari (2001) – however their findings were based on bibliographic metadata. In the case of Web IR Tombros et al. (2003; 2005) similarly found that content and layout-related features accounted for most criteria applied in relevance assessments and no explicit situational statements appeared, except for utterances like: "this is hot, man". This kind of higher order relevance rather collapses or disintegrates into lower order relevance, like topicality or algorithmic relevance.

This phenomenon resembles the so-called cognitive "free fall" discussion in the cognitive theory for IR and information seeking (Ingwersen and Järvelin, 2005, p. 33-38): that conveyed messages loses their meaning and disintegrate into the morpho-lexical linguistic level of communication.

The issue here is thus quite simple: it may be a certain combination of accessible features viewed at a certain point during interaction which, in a personalized way, triggers the situational assessment. Operationally however, only the topical and pertinent relevance features directly related to documents and contents, like layout, date, place, names, images, etc., or added task related keys, remain as evidence during a session (and evaluation experiment).

These phenomena of disintegration of meaning during communication and collapse of higher order (situational) relevance into lower order relevance exactly form the bridges that bring together the Laboratory and Cognitive IR Frameworks: on the one hand one hoped for sophisticated answers to research questions during evaluations but must accept rather simplistic evidences of interactive processes, judgments and behavior on both system and searcher side; but on the other hand one may extend the investigations into the field, out into the context in a controlled and constructive manner, Section 4.

4 DRIFTING OUTSIDE THE LABORATORY CAVE AND INTO CONTEXT

When moving out into the contextual parameters surrounding the cave of the Laboratory Framework it is mandatory deliberately to apply a firm hand on selected variables, as proposed by the Cognitive Research Framework (Ingwersen and Järvelin, 2005, p. 313-376), and illustrated by the case below. In total, the Cognitive Framework for IR involves 9 dimensions of variables, each with their own values. The dimensions derive from the five components *and* the interaction processes of the holistic cognitive framework, Figure 5, with the actor defining three dimensions and the social context two. One might call this scenario Interactive IR Evaluation 'Light' – Figure 7.

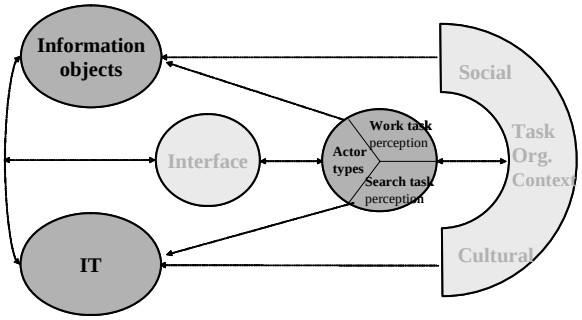

Figure 7. Dimensions from the holistic Cognitive Framework included in IR Evaluation Light. (Ingwersen and Järvelin, 2005, p. 364).

4.1 Research Question and Types of Variables

This sample research setting incorporates the basic laboratory model components, but extends it by including the seeking actor into an interactive (IR) scenario, Figure 7, that is, independent variables from three dimensions of the framework. The actor is seen in the light of three main dimensions of variables, some of which are controlled, neutralized or independent, depending on the research question:

- Actor type variables
- Perceived work task variables

- Perceived search task variables

The research question could be:

Given a specific organizational context X with known work task types, which IR method performs best considering different searcher work task experiences and knowledge and a variation of document types?

The context might, for instance, be a selected medical domain and organizational environment. Typical work task types are clinical diagnosis, treatment, clinical testing, surgical procedures and execution, medical prescriptions, etc. The matching techniques undergoing performance evaluation are, e.g., a probabilistic model versus a browsing based access tool. The searching actors are either experienced doctors vs. 1st year medical students. The documents used as knowledge sources are either academic full-text journal articles, or academic web sites. The searches to be done during experimentation are instigated by a set of realistic simulated work task situations given to the test persons (Borlund, 2003b). The set is chosen to be of the semantically closed kind, but could also consist of naturalistic work task assignments lacking cover stories. Preferably, such cover stories / assignments should lead to search tasks adhering to the factual Information Need Type. Cover stories or assignments might consist of Roentgen photos or video shots by micro cameras of specific cases – largely replacing written statements. The actors may execute their information access as they would like in realistic terms, but cannot make use of human information sources (Ingwersen and Järvelin, 2005, p. 365-366).

Table 1. Independent variables (dark shading, framed) and controlled variables (light shading) combined in an IIR experiment.

Document and Source types	Algorithmic IT Component	Actor Characteristics	Perceived Work Task	Perceived Search Task
Doc. Structure	Doc. Metadata rep.	Domain Knowledge	Struture/Openness	Inform. Need Types
Doc. Types	Doc. Content rep.	IS&R Knowledge	Strategy/Practice	Structure/Type
Doc. Genres	Doc. Structural rep.	Work Task Exp.	Granularity/Size	Strategy/Practice
Information Types	Req. Metadata rep.	Search Task Exp.	Dependencies	Complexity/Specific.
Comm. Function.	Req. Content rep.	Work Task Stage	Requirements	Dependencies
Sign Language	Req. Structural rep.	Context Perception	Domains/Context	Stability
Layout & Style	Match Methods	Constraints	…	Domains/Context
Doc. Isness	…	Motivat./Emotion		…
Link Structures		…		
Human Source				
…				

The motivation for the research is the assumption that the traditional academic documents are better information sources for solving the work tasks by the experienced doctors than web-based material. Secondly, it is interesting to find out which access technique, the browsing based technique or the probabilistic engine, is more effective.

4.2 Experimental Setting

Table 1 demonstrates which variables (dark shaded framed cells) from the three central research dimensions that are involved in answering the research question. Each variable in question may take a range of values. For instance, in general the Document Dimension variable Document Type contains values ranging from newspapers over monographs to journal articles, conference papers, music recordings, Web-based data, etc. In the specific case the range has been limited to a few selected types, as stated above. In this research question the Work Task Structure/Openness, the Domain/Context as well as the Information Need Type and Human Source variables are all controlled (lightly shaded cells) – since all the simulated work task situations are of the factual type and from a selected domain. By being the same throughout the investigations the Interface Component as well as the Socio-Organizational Context dimensions are also controlled. The dependent variable is performance. All other variables (white background) suggest potential hidden variables.

The proposed research design operates with combining the selected variables from three dimensions in such a way that, e.g., 32 test persons (16 doctors and 16 medical students), 8 simulated work tasks / assignments (Q1-Q8), the two retrieval methods (a and b), and the two document types (D1 and D2) are systematically and symmetrically combined during the investigation. The design implies that 8 test persons (doctors) as well as 8 test students each search two assignments (Q1-2) for method (a) + document type (D1) *and* (Q3-4) for method (a) + D2. The same test person groups then switch to search (Q5-6) and (Q7-8) via the method (b) + D1/D2 configurations respectively. Eight new test doctors and eight new medical students then repeat the design symmetrically, so that the assignments (Q1-4) are tested on the two configurations: method (b) + D1/D2 and (Q5-8) are tested on method (a) + D1/D2. The operations can be done by means of contingency tables.

The proposed research design thus operates with eight assignments per test person, a doable set of

search tasks, and 32 search events defined by the four assignments dealing with each model/document type combination – in total 64 search events over all eight assignments for each combination. Hence, for each searcher type there are generated 32 search events per combination. In total 256 searches (32 persons x 8 assignments) are conducted. The design makes it possible also to study the searcher behavior of the different groups. Obviously, if it is not feasible to reach the necessary number of test persons, each participant is then required do more than eight searches. Then the behavioral aspects of the investigation become less statistically reliable. The assignments do not have to be carried out during one day, but can be distributed over several days.

5 POINTS OF SUMMARIZATION FOR DISCUSSION AND CONCLUSION

The holistic Cognitive Framework as well as its nine-dimensional Research Framework offers broader and more profound conceptual explanations as well as hypotheses than the Laboratory Framework or the various user-centered approaches on IR in isolation. One may advance the points that are the most central for our deeper understanding of information retrieval and the study of the phenomena associated with IR:

- Real users do not necessarily have ready made verbose needs;
- Real users interpret their situations differently – even if put into the "same" situation (as far as that is possible, e.g., by means of simulated task situations);
- Even if experienced professionals, they may approach the information problems from different angles;
- Real users therefore construct quite different queries and are for long known to assess document relevance differently (Cleverdon, 1984)
- Real users face vocabulary problems and do not know the collection well; therefore initial queries may fail badly and (pseudo) RF may not work.
- Therefore they may issue many consecutive queries on the same (but evolving) topic/need as they learn on the fly what works and what does not.
- Consequently, while ranking well for any given query is important, it is at least equally important to help the user to arrive at a good query (and if it is really good, any ranking method works well)

The Research Framework suggests how to deal with them:

- Application of research designs like that in Section 4 above or other settings combining different variables from the nine dimensions of the Cognitive Framework (Ingwersen and Järvelin, 2005, p. 359-380). The Framework suggests to involve maximum three independent variables for reasons of complexity;
- Comparing retrieval in different types of collections;
- Comparing experts and novices – like in the case above, Section 4;
- Comparing natural work tasks with simulated tasks which have been consciously manipulated to be more or less vaguely designed. The natural tasks serve as the control mechanism during the (field) experiments.

We have analyzed and discussed the Laboratory Research Framework and shown how it fits into the holistic Cognitive Framework and the consequential Cognitive Research Framework. We have not simply outlined central problems and issues to be solved for IR research but also pointed to their solutions based on theoretical foundations and true research models. Most importantly, we have demonstrated how the Cognitive Framework may lead to experimental designs that are realistic but, at the same time, also controlled. The Framework thus attempts to incorporate the best features from the Laboratory Framework and its algorithmic models, at the same time as it points to new metrics for new experimental situations: IR in context.

It is only via experimentation in the laboratories *and* the field that we obtain an understanding of the best retrieval steps of algorithmic and behavioral nature. It is similarly vital to find good handles in the interactive processes, and how and when people realize their existence. Interface design and evaluation hence becomes more central to IR research because the implicit or explicit RF and types of query modification require interfaces that accommodate the searcher *and* the system in a balanced way. Further, detecting ways to support good (initial) queries in context is invaluable for IR system development. Single query based

experimentation and metrics do not help to identify which interactive handles and steps that are most effective. The evolution of research frameworks in IR makes it possible to move out into an open landscape of investigation.

Acknowledgements

The authors wish to thank the Nordic Research School of Library and Information Science (NORSLIS) for travel support.

REFERENCES

Barry, C.L., Schamber, L., 1998. Users' criteria for relevance evaluation: A cross-situational comparison. *Information Processing and Management, 31*(2/3): 219-236.

Bookstein, A., Swanson, D., 1974. Probabilistic models for automatic indexing. *Journal of the American Society for Information Science, 25*(5): 312-318.

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

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

Borlund, P. (2003b). The IIR evaluation model: A framework for evaluation of interactive information retrieval systems. *Information Research, 8*(3), paper no. 152 [Available at: http://informationr.net/ir/8-3/paper152.html. Cited May 13, 2003.]

Borlund, P., Ingwersen, P., 1998. Measures of relative relevance and ranked half-life: Performance indicators for interactive IR. In: W.B. Croft, et al. (Eds.), 21st ACM-SIGIR Conference. ACM Press: 324–331.

Bunge, M., 1967. *Scientific Research*. Springer. Heidelberg.

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

Cleverdon, C.W., 1984. Optimizing convenient online access to bibliographic databases. *Information services and Use*, 4: 37-47.

Cosijn, E., Ingwersen, P., 2000. Dimensions of relevance. *Information Processing and Management, 36*: 533-550.

Egghe, L., Rousseau, R., 1990. *Introduction to Informetrics: Quantitative Methods in Library, Documentation and Information Science*. Elsevier. Amsterdam.

Engelbart, D., 1962. Augmenting Human Intellect: A Conceptual Framework: Stanford Research Institute. Menlo Park, CA.

Ingwersen, P. (1992). *Information Retrieval Interaction.* Taylor Graham. London

Ingwersen, P. (1996). Cognitive perspectives of information retrieval interaction: Elements of a cognitive IR theory. *Journal of Documentation, 52*(1): 3-50.

Ingwersen, P., Järvelin, K., 2005. *The Turn: Integration of Information Seeking and Retrieval in Context*. Springer. Heidelberg

Järvelin, K., 2007. An analysis of two approaches in information retrieval: from frameworks to study designs. *Journal of American Society for Information Science and Technology*, 58(7), 971-986.

Järvelin, K., Kekäläinen, J., 2000. IR evaluation methods for retrieving highly relevant documents. In: 23rd ACM-SIGIR Conference. ACM Press: 41-48.

Järvelin, K., Kekäläinen, J., 2002. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (ACM TOIS), 20*(4): 422-446.

Katz, S., 1996. Distribution of content words and phrases in text and language modelling. *Natural Language Engineering, 2*(1): 15-60.

Kekäläinen, J. (2005). Binary and graded relevance in IR evaluations - Comparison of the effects on ranking of IR systems. *Information Processing and Management, 41*(5), 1019-1033.

Kekäläinen, J. Järvelin, K., 1998. The impact of query structure and query expansion on retrieval performance. In: Croft, W.B. et al. (Eds.), 21st ACM-SIGIR Conference. ACM Press: 130-137.

Kekäläinen, J., Järvelin, K., 2002a. Evaluating information retrieval systems under the challenges of interaction and multi-dimensional dynamic relevance. In: Bruce, H. et al. (Eds.), The CoLIS 4 Conference: 253-270.

Kekäläinen, J., Järvelin, K. 2002b. Using graded relevance assessments in IR evaluation. *Journal of the American Society for Information Science and Technology, 53*(13): 1120-1129.

Keskustalo, H., Järvelin, K., Pirkola, A. (2006). The Effects of Relevance Feedback Quality and Quantity in Interactive Relevance Feedback: A Simulation Based on User Modeling. In: Lalmas, M. and al. (Eds.), 28th European Conference on Information Retrieval ECIR'06, London, April 2006. Heidelberg: Springer, Lecture Notes in Computer Science vol. 3936, pp. 191-204.

Larsen, B., Ingwersen, P., Kekäläinen, J., 2006. The polyrepresentation continuum in IR. In: Ruthven, I. et al. (eds.), *Information Interaction in Context: IIiX*. Royal School of LIS, Copenhagen: 148-162.

Magennis, M., van Rijsbergen, C.J., 1997. The potential and actual effectiveness of interactive query expansion. In: Belkin, N. J., Narasimhalu, A. D. and Willett, P. (Eds.), 20th ACM-SIGIR Conference. ACM Press: 324-332.

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

Robertson, S.E., Hancock-Beaulieu, M., 1992. On the evaluation of IR systems. *Information Processing and Management, 28*(4): 219-236.

Saracevic, T., 1996. Relevance reconsidered '96. In: Ingwersen, P. and Pors, N.O. (Eds.), 2nd CoLIS

Conference. Royal School of LIS, Copenhagen: 201-218.

Skov, M., Larsen, B., Ingwersen, P., 2006. Inter and intra-document contexts applied in polyrepresentation. In: Ruthven, I. et al. (eds.), *Information Interaction in Context: IIiX*. Royal School of LIS, Copenhagen: 163-170.

Sormunen, E., 2002. Liberal relevance criteria of TREC – Counting on negligible documents? In: Beaulieu, M. et al (Eds.), 25th ACM-SIGIR Conference: 320-330.

Tombros, A., Ruthven, I., Jose, J., 2003. Searchers' criteria for assessing web pages. Proceedings of the 26th ACM-SIGIR Conference, ACM-Press: 385-386.

Tombros, A., Ruthven, I., Jose, J.M., 2005. How users access web pages for information seeking. *Journal of the American Society for Information Science and Technology, 56*(4): 327-344.

Vakkari, P., 2001. Changes in search tactics and relevance judgments in preparing a research proposal: A summary of findings of a longitudinal study. *Information Retrieval, 4*(3/4): 295-310.

Vorhees, E.M., 2001. Evaluation by highly relevant documents. In: 24th ACM-SIGIR Conference. ACM Press: 74-82.

Voorhees, E.M., 1998. Variations in relevance judgments and the measurement of retrieval effectiveness. In: Croft, W.B. et al. (Eds.), 21st ACM-SIGIR Conference. ACM Press: 315-323.

Wang, A.P., de Vries, A.P., Reinders, M.J.T., 2006. In Lalmas, M., Tombros, A. (eds), Proceedings of the Annual European Conference on Information Retrieval (ECIR): 37-48.

White, R. W., 2006. Using searcher simulations to redesign a polyrepresentative implicit feedback interface. *Information Processing and Management, 42*(5): 1185-1202.

White, R., Ruthven, I., José, J.M., van Rijsbergen, C.J., 2005. Evaluating implicit feedback models using searcher simulations. ACM Transactions on Information Systems, 23(3): 225-361.

Willett, P., 1988. Recent trends in hierarchic document clustering: A critic review. *Information Processing and Management, 24*(5): 577-597.