IR Research: Systems, Interaction, Evaluation and Theories[1]

Kalervo Järvelin
School of Information Sciences
University of Tampere, Finland
kalervo.jarvelin@uta.fi

Abstract

The practical goal of information retrieval (IR) research is to create ways to support humans to better access information in order to better carry out their tasks. Because of this, IR research has a primarily technological interest in knowledge creation – how to interact with information (better)? IR research therefore has a constructive aspect (to create novel systems) and an evaluative aspect (are they any good?). Evaluation of IR effectiveness is guided by a theory on factors that affect effectiveness. True science is about theory development, i.e., understanding and explaining, making hypotheses and testing them. Theories express structured explanatory relationships between variables such as "type of document indexing" and "quality of ranking measured by MAP". Theories are the better, the wider range of phenomena they are able cover accurately. The paper argues that most existing theories of IR are focused on a narrow scope, theories of ranking. To fulfill its task of supporting human information access, theories that go beyond the evaluation of ranking are highly desired but face many challenges. We discuss three additional types of IR theories: theories of searching, theories of information access, and theories of information interaction.

1 Introduction

The practical goal of information retrieval (IR) research is to create ways to support humans to better access information in order to better carry out their (work) tasks. Because of this, IR research has a primarily technological interest in knowledge creation – how to find information (better)? IR research therefore has a constructive aspect (to create novel systems) and an evaluative aspect (are they any good?). Evaluation requires some object that is evaluated and some goal that should be achieved or served. Evaluation of IR effectiveness is guided by a theory on factors that affect effectiveness.

Practical life with all its variability is difficult and expensive to investigate. Therefore surrogate and more easily measurable goals have been employed in IR evaluation, typically the quality of the

[1] http://www.info.uta.fi/tutkimus/fire/archive/2011/ECIR'11-Keynote-slides-Kal.pdf

ranked result list instead of the work task result. The task performance process may also be cut down from a work task to a search task and down to running an individual query in a test collection. This simplification has led to standardization of research designs and tremendous success in IR research. However, as the goals and systems drift farther away from the practical life condition, one need to ask, whether the findings still best serve the initial goal of evaluation, supporting human task performance?

It is important to evaluate all subsystems of information retrieval processes, in addition to the search engines. Through a wider perspective one may be able to put the subsystems and their contributions in relation with each other. We will therefore discuss nested IR evaluation frameworks ranging from IR system centered evaluation to work-task based evaluation. We will also point to the Pandora's box of problems that the enlargement of the scope of research entails. Is science at risk here?

The contributions of a research area, in addition to constructive and evaluative contributions, may be generally empirical, theoretical and methodological. Why should anyone in IR care about anything beyond IR experimentation (i.e. evaluation) using test collections? The Cranfield model seeks to relate texts (documents), queries, their representations and matching to topical relevance in ranked output. Who relates this, and a range of possible other contributing factors, to outcomes in search task performance or work task performance? We discuss three additional types of IR theories: theories of searching, theories of information access, and theories of information interaction. They may help clarify IR research and evaluation designs and focus research efforts on technical issues and evaluation.

The paper is organized as follows: We begin with a few words about the goals and context of IR. This leads, Section 3, to a discussion of the aims in IR research, followed by some words on interaction in Section 4. Evaluation, Section 5, is a key aspect in IR research, but there are many kinds of IR evaluation. We shall discuss the objects that are evaluated as well as the goals served. Finally in Section 6, evaluation contributes to, and benefits from, understanding which factors explain human task performance or information interaction, that is, theories.

2 On Information Retrieval out There

The goal of information retrieval in practical life is to access information in order to better carry out human tasks. Such primary tasks may be work-related or leisurely. Serving such tasks causes that IR in practice is of *instrumental* value.

In real life, information access is often a necessity for task performance, but finding useful information for task performance is often a challenge. Therefore in information-intensive tasks, tools are used to augment insufficient human capabilities. Tools shape the tasks and tools should be designed or adapted for the tasks. How well do we understand information-intensive tasks and the position of IR in their augmentation?

Some 50 years ago, D. C. Engelbart [10], the inventor of the computer mouse, proposed a framework for augmenting human intellect. Figure 1 is a bit modified version of this framework, suggesting that augmentation of work tasks in context may take place through tools, methods, knowledge, and training. The goals of information retrieval in practice must be related to such augmentation, so a proper question is how information retrieval is related to augmenting human intellect.

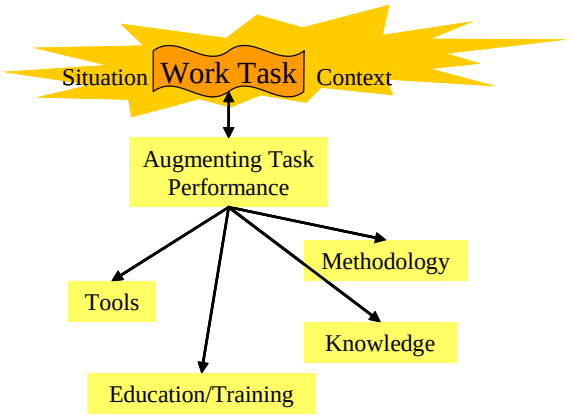

Fig. 1. A framework for augmenting human intellect (modified from [10])

Figure 2 provides an example focusing on augmenting the task performer's work methodology and relevant knowledge. Information may be accessed through various channels, created or remembered. The figure suggests a means-ends hierarchy where some of the paths lead to information retrieval. Therefore the figure suggests that, in principle, IR is one means of information access in the task performer's personal information ecology where the end is augmentation. Multiple channels and sources are often used concurrently. Indeed, in real life this seems to be the case: for a single work task, multiple channels and sources may be harnessed for information access (Figure 3).

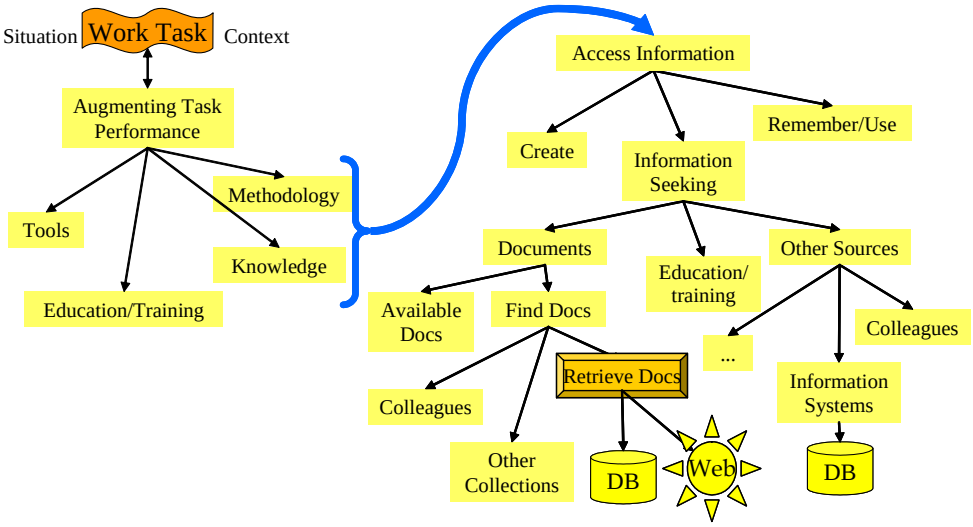

Fig. 2. A sample means-ends hierarchy connecting work task augmentation and information retrieval [15]

Figure 3 presents the step-by-step performance of a semi-complex work task session consisting of 28 steps over a duration of more than two hours. The task was performed in a biotechnology research unit. The top line indicates the five queries employed. The bottom line approximates task duration in chunks of 15 minutes. Note that they are, graphically, of varying lengths because some time-slots involve many events while others may involve just examining a found resource. The shaded column is a 15-minute break.

The middle area represents the structure of the information ecology of the researchers and is divided into broad horizontal channels: search engines, web sites, literature databases, bio-databases, the PC

tools, and other. Each track within a channel represents the same unique resource throughout the session while different lines indicate different resources. The solid arrows represent the workflow, the dashed arrows represent data flows, and the dotted arrows represent transitions by links.

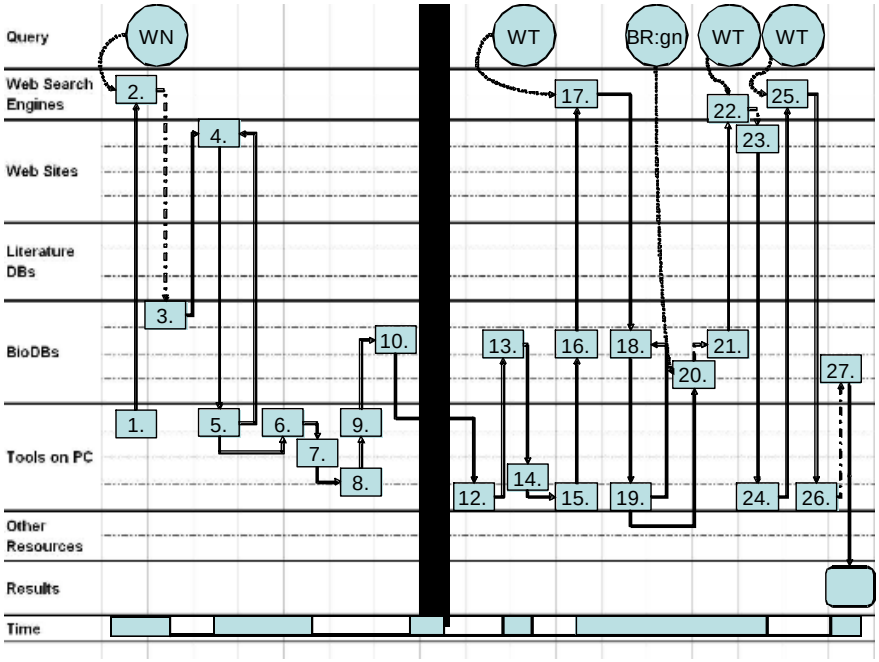

Fig. 3. A sample work task process in Molecular Medicine [19]

While the majority of action takes place between three bio resources in this task, this session spreads across four channels. Overall, we note that the task progresses through several tools in several channels and may return to a given tool after some visits to other tools. The task performer harnesses and integrates the tools for the task as is most natural and convenient for him or her. This depends on his/her personal information ecology. We believe that we can and will find similar features in many other task domains, not just in molecular medicine and not just in research contexts.

A task performer's *personal information ecology* consists of the *objective* and *subjective* aspects of channels, tools, services, documents, or resources that a person can use for information access and organization. The objective aspects relate to what really is available while the subjective aspects relate to *perceived* availability. The latter obviously works in the bounds of the former but they are not equal. The latter is more important because the task performer uses it for information access. It tells situationally, which ecological niches are valuable. In a subjective information ecology, various resources are *seen to serve* varying ends under varying conditions like ease of use, or costs. Resources that are useful fit to some *info-ecological niche* of the task performer.

The design implication is that resources should be designed for their niches, typical uses and coordinated use. When designing IR systems -- how often do we think about their ecology? Much of the IR literature reflects a design approach that assumes that IR systems are separate systems, or even the only systems used, and perceived as such by their users. This view seems far from truth: as Figure 3 suggests, several kinds of information systems may be used in an interleaved fashion. As a community, we don't know how the systems are really used.

3 On Information Retrieval Research

The goals of IR research can be stated in many ways (see e.g. [15]). Here is one more:

> The ultimate goal of IR research is to *create ways to support* humans to better access information in order to better carry out their tasks.

Such an instrumental goal makes IR research a branch of technology rather than science. Consequently, IR research has a primarily technical interest in knowledge creation – how to find information (better)?

There are two aspects in our research - a constructive one (to create novel systems) and an evaluation aspect (are they any good?). We tend to be proud if we can state at conferences that "*My brand new search engine is better than yours!"* Let us look at the engineering aspect first.

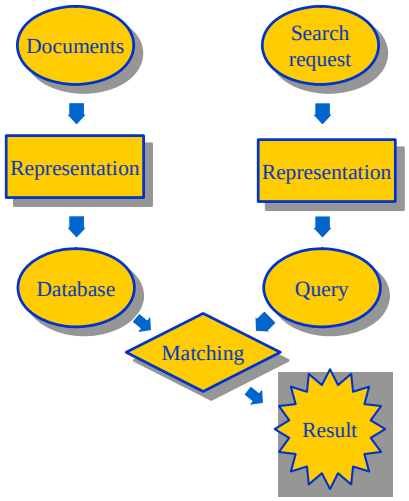

Fig. 4. IR engineering focus

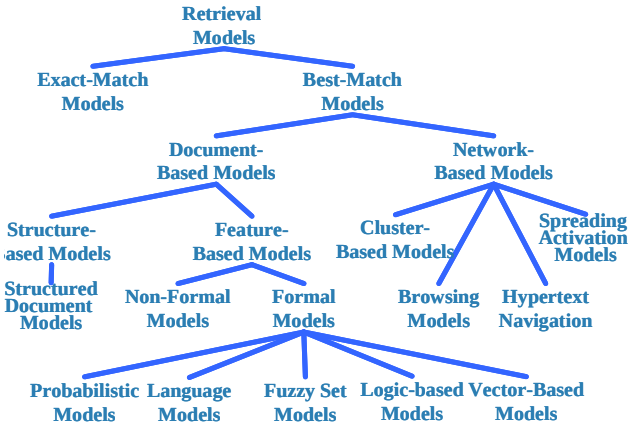

Fig. 5. IR models ([15], modified from [6])

Figure 4 presents the Cranfield IR engineering framework with documents, requests, their representations, their matching and the results. A lot of effort has been expended on the representations and matching when designing search engines. In particular, the history of IR research has developed a range of retrieval models, some of which are organized into a tree in Figure 5. Often in IR literature, when one discusses IR theory, such IR models are given as examples of IR theory.

The engineering approach is necessary – without the engines the IR community might not exist at all. Still, this is not all that is needed. Comparing to automobile industry, such a focus would mean spending the engineering efforts on the engine, transmission and wheels with little attention to comfort, safety or navigation. Returning to search engines, the foci in Figures 4 and 5 do not offer much regarding the analysis of human interaction with search engines, databases or information.

4 Interaction in Information Retrieval

Searchers interact with search engines, because they have to or do not know for the moment a better means in their information ecology. The study of, and design for, IR interaction is becoming increasingly important in IR research. It is widely acknowledged that the practice of information

retrieval means no single shots at the database but interaction with search engines, databases and information. As behavior this shows as query reformulation, browsing and examination of documents and their snippets. The Cranfield approach to IR and the engineering focus are not most helpful here.

There is a long tradition of IR interaction studies in Information Science. One may think of the work by Bates [4], Borgman [8], Fidel [11], Saracevic [22], for example on searching styles, strategies, tactics, and interfaces. Its value may be shadowed by the fact that many of the studies pre-exist the TREC era and because the context was professional searchers and Boolean retrieval systems. In the 80's there was much interest in expert systems in IR interaction - in cognitive viewpoint, remember Belkin [5] [6], Ingwersen [14] and others. In the 90's the community's focus however shifted to scaling up, ranking results in large collections, the web, etc. Recent years have brought interaction back to focus with a stronger understanding that searchers don't only launch one query in a session but browse and reformulate, that is, interact with the engine. There are new conscious efforts for accounting for interaction is system design and evaluation. The searchers' real needs however are to interact with information, not the engines per se, because they need to construct new knowledge.

One popular approach to the analysis of interaction is log analysis. We may wish to observe in a search engine's query log sequences of queries or entire sessions (Figure 6). We may then analyze query reformulations and the ranks of links the searchers click, dwell times, etc. We may count the number of keywords and links clicked. We may then try predicting the information needs or search goals. This works often well, but may sometimes give a biased view especially if there is a larger task underlying the system use. This may be understood when looking at what takes place at the user side.

Fig. 6. A simplified query log representation with gaps between queries

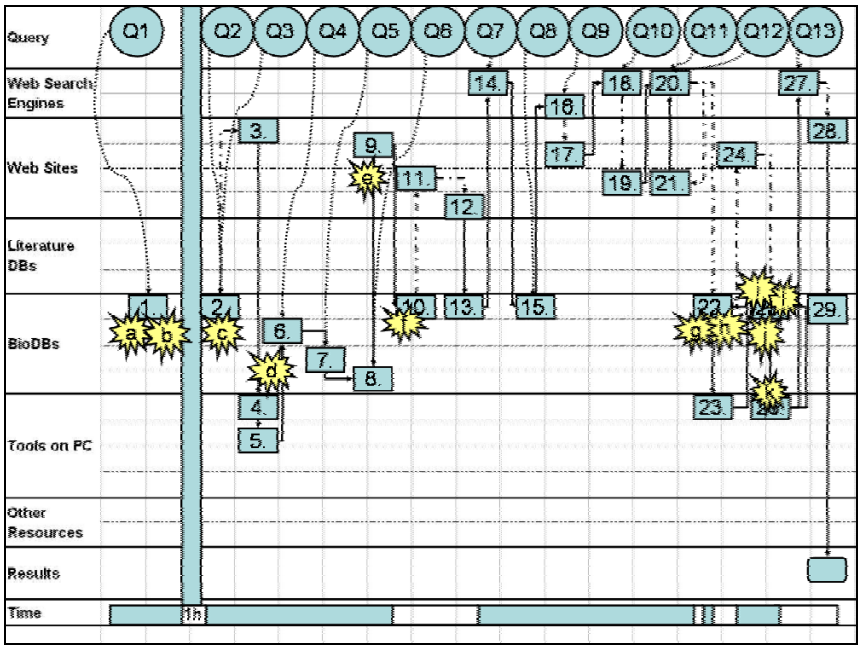

Fig. 7. Multiple queries Q1-Q13 for multiple engines of various types in task steps 1-29 with barriers a-l

Figure 6 shows seven queries from a single IP address in a single search engines search log. The entire sequence might meet some popular session criterion (e.g., all launched within 15 minutes) and be therefore considered as a session and analyzed for its goals. However, this single server-side log cannot show what else the searcher may have been doing. Figure 7 represents a real life task session with 13 queries in several types of search engines. The notations are Figure 7 are the same as in Figure 3 but for the barriers a-l (the flash symbols) the searcher encounters in various steps of the task (see [20]). The real need is to interact with information to construct some new knowledge, and this is not a single source - single query activity. Search engines may only aid in this - more or less.

5 Evaluation in Information Retrieval Research

Evaluation is often referred to as a hallmark and distinctive feature of IR research. No claim in IR is granted any merit unless it is shown, through rigorous evaluation, that the claim is well founded. Technological innovation is the driving force in the transformation of information access but alone is not sufficient for progress. Innovations must also be show as useful.

Evaluation, in general, is the systematic determination of merit and significance of something using criteria against some standards.[2] Evaluation requires an object that is evaluated and some goal that should be achieved or served. In IR, both can be set in many ways. The object usually is an IR system or a system component – but could be as broad as a human performing a task. The goal is often the quality of the retrieved result – but what is the retrieved result and what its quality? There are alternative answers, which lead to different kinds of IR evaluation.

Practical life information access is difficult and expensive to investigate due to its variability. Therefore surrogate and more easily measurable goals are employed in IR evaluation, typically the quality of the ranking of the result instead of the work task outcome. The system often is reduced from work task performance to search task performance, or even down to running a query in a test collection.

Such simplification has offered great standardization of research designs and led to tremendous success in IR research. This is based on the shared test data and thus comparable test results. In principle, this would lead to cumulating knowledge and annually improving retrieval technology.

This positive outlook however has been challenged by Armstrong and colleagues [3]. They argue that too weak baselines have continuously been used in IR experiments for achieving statistically significant – and thus reportable – research results. Therefore they claim that the reported improvements in performance don't add up.

This claimed lack of progress might be remedied through improved practices in IR evaluation. For example, one could carefully employ harder baselines. Thus one could identify true progress in retrieval effectiveness when there is some. Should one then be satisfied with 70% MAP, 90% MAP or 95% MAP as evaluation results? How far would one need to push the evaluation results?

The Cranfield model of IR evaluation (Figure 8) guides us to aim at 100% MAP or nDCG as the evaluation result. Would that make the designer happy? An unbeatable search engine? The IR problem solved and unemployment for all others as the outlook?

[2] See Wikipedia, http://en.wikipedia.org/wiki/Evaluation

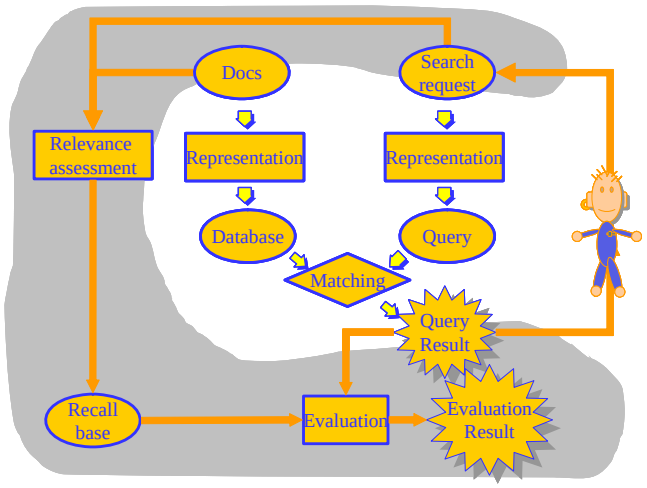

Fig. 8. The Cranfield model of IR experimentation (e.g. [15])

I think that we need an evaluation framework that incorporates, say, a human work task, its evaluation and seeks to establish a correlation of our standard output measures such as MAP or nDCG with some outcome measure. Figure 9 gives, in addition to the Cranfield style test collection components, also a human task with its outcome that is evaluated for its quality. Now a crucial question is the correlation of the search evaluation result and the task evaluation result. Such a correlation would justify both our IR experiments and the pursuit of better ranking. More of MAP would bring better task outcomes. This is what we do in IR, but many of us turn hesitant, if this issue is made explicit.

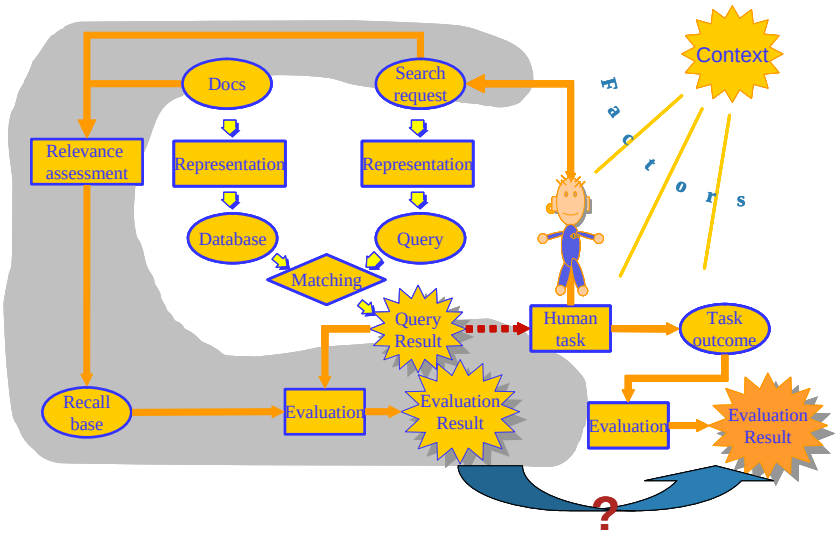

Fig. 9. Evaluation by task outcomes

Unfortunately, such a strong correlation is lacking. Several recent studies have elaborated on this – and argued that there is only a weak connection between IR performance and human task performance (see e.g. [1] [13] [23] [24]). This seems to hold even if the human task is very close to retrieval – such as finding one or a few relevant documents. A recent radical example [26] argues based on empirical findings that search interaction effort degrades search output quality but improves task outcome quality. If this would be generally true, there would be a correlation between the two –

but a negative one.

When claiming that the correlation of search and task evaluation results is not strong, I'm not suggesting that random results are as good as ranked ones. Some level of quality is certainly needed. However, if the findings referred to above are true there are some consequences regarding the use of Cranfield style evaluating results [17]:

- there is the risk of recommending a system that may not be optimal for humans;
- the risk is not likely to change, no matter how much we work with the retrieval models and current Cranfield style output evaluation methods.

In order to break this circle, the community has begun to look into interactive IR experiments.

Bringing the user properly into IR evaluation is already a challenge: we are struggling with experimental methods that account for search strategies, interaction, searcher characteristics, for example. However, if the evaluation target is improved output in the traditional sense (say MAP), the target is insufficient. *More of MAP does not guarantee more of task outcome*. A greater challenge is to find which factors correlate with "more of task outcome". Some of them may be close to IR systems and interaction.

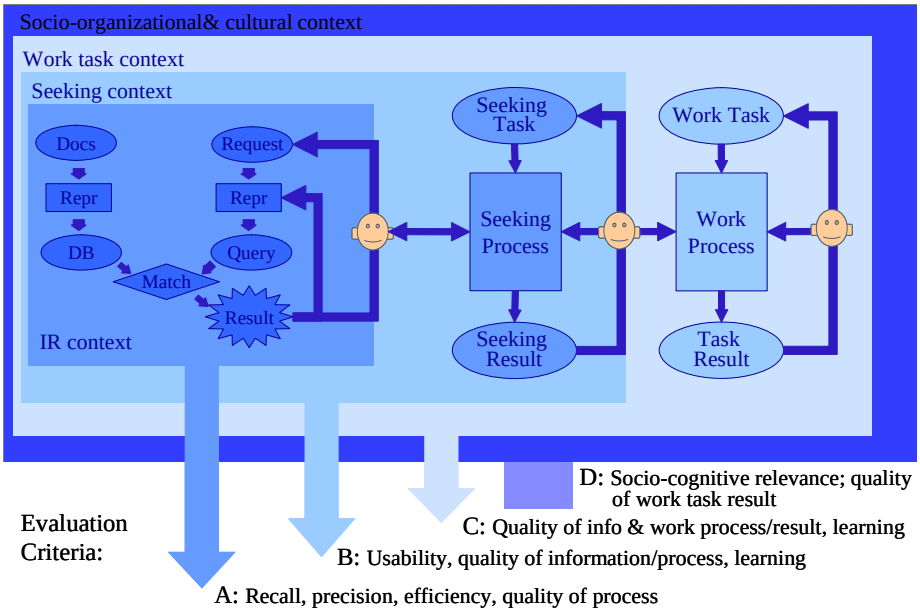

Fig. 10. Nested evaluation frameworks ([15] [18])

We may analyze the evaluation problem through the nested evaluation frameworks in Figure 10: the IR context, seeking context, work-task context, and socio-organizational and cultural contexts. They return the ultimate goal of IR to the scene, the human task performance. Sample evaluation criteria in each context are given at A – D. First, IR may be designed and evaluated in its own specific context – the test-collection approach (A) using traditional output evaluation measures and the quality of the search process like searcher's effort (time) and satisfaction. Because IR belongs to the searcher's information seeking context (B), it is one means of access (cf. Figure 2). This context provides a variety of information sources and systems for consultation. One may evaluate what is the contribution of an IR system at the end of a seeking process. Next, the real impact of information retrieval is its contribution to the work task (e.g., effort, time) and the quality of the task result (the criteria C). Finally, work tasks are performed in a socio-organizational and cultural context (D) and

may be evaluated for their contribution therein.

In addition to asking whether one obtains an excellent ranking from a search engine, one may ask other questions as well:

- Is the information accessed through the interaction better? Did the searcher learn something? Which factors contribute to that?
- Does the information retrieved contribute to the quality of the work task result? Which factors contribute to that?

If we do not look at such issues we run the risk of suboptimization of IR systems and cannot convincingly claim that IR is a useful exercise. It may be, but we are not sure.

One might criticize me by saying that the effects of ranked outputs are mixed with other effects so that discerning the contribution of an IR system from all the other factors in such wider contexts is difficult. Therefore one wouldn't be able to measure improvements in search engines reliably. We would no more be able to argue rationally, whose search engine is the best. This is true. However, I would still like to turn this claim upside down: If the evaluated output of an IR system does not show in task performance, why so much fuss about IR experiments and their evaluation metrics? Shouldn't we focus on such factors that apparently have an effect on, say, task performance? Can we even today argue rationally, whose search engine is the best for real life applications?

6 Beyond Evaluation: Theories

It would be simplistic to stop with evaluation in IR research. True science is about theory development, i.e., understanding and explaining, making hypotheses and testing them. Theories are the better, the wider range of phenomena they are able cover accurately. Theories and their constituents may be characterized briefly as follows (see [16] for a longer discussion):

- Typically theories are seen as to consist of systematic collections of theoretical and empirical laws. A theory explains observed regularities and hypothesizes novel ones. Further, a theory provides deeper understanding of phenomena by using theoretical concepts that go beyond immediate observations. Scientific theories represent reality, and guide research by suggesting novel hypotheses.
- Scientific laws are typically classified as empirical or theoretical. Empirical laws express verified relationships between observable objects, properties or events (variables). Theoretical laws refer to directly non-observable objects or properties. Laws may express deterministic or probabilistic regularities (e.g. Zipf's Law).
- A hypothesis states a verifiable fact, e.g. a relationship between some variables, whose truth is unknown. A hypothesis needs to explain the fact and be testable. A hypothesis is accepted if it passes a rigorous test.
- In a study design, one may specify the interaction of several types of variables to be examined: *dependent* variables – the variation of which is explained; *independent* variables – the ones systematically varied in order to see the responses in the dependent ones; *concomitant* variables – the ones fixed to prevent uncontrolled variation in the results.

Theory development is about variables and their variation. Some are explained, some others are used to explain the variation of the former or control unwanted variation. Theories of IR may be analyzed and classified based on their constituent variables. We shall use the contexts in Figure 10 as a discussion tool for theories of IR below.

6.1 Theory of Ranking

The most popular theory of IR is the theory of ranking, Figure 11. This theory is clearly needed and also pretty well understood. It tries to explain the variation of the quality of ranking, measured in terms of some effectiveness metrics (e.g. MAP), mainly through the variation of document and query representations and their matching models and (often) through the control of documents and requests (using test collections).

When one asks around for theories in Information Retrieval, a typical answer is that the IR Models (see Figure 5) are the IR theories. This suggests that we have a probabilistic theory of IR, a language model theory of IR, etc. These theories relate document and query representations, matching, and the quality of the output. The engineering aspect works well, but the prediction of the quality of output remains often quite uncertain. In other words, given queries, a collection, and the representation and matching methods, these theories *determine* what one gets as output. Whether that is any good, remains *more uncertain*. As Robertson put it in his Salton Award lecture at SIGIR 2000 [21], in IR we may gain theoretical insight but no strong predictions through experimentation.

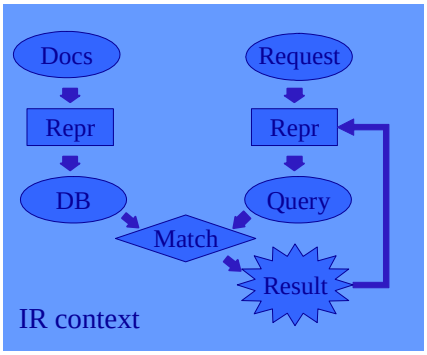

Fig. 11. Components of the theory of ranking

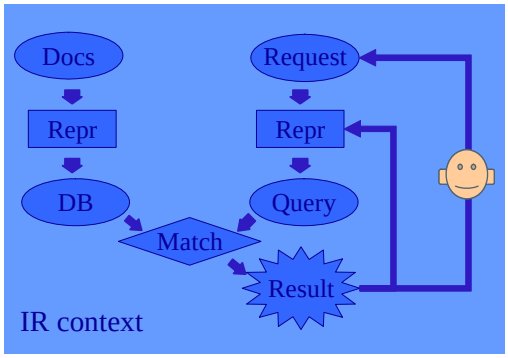

Fig. 12. Components of the theory of searching

6.2 Theory of Searching

Another theory class within the IR context adds a human searcher to the study designs (Figure 12). In campaigns such as the interactive TREC[3], a human being acts as an operator of search engines in order to see whether the results of the theories of ranking can be replicated when a human is present a the steering wheel. It was found early that searchers and topics generally have a greater effect on search results than the IR technique [12]. Adding variables representing the human operator (e.g. cognitive skills) or the process (e.g. time allocated for searching) is an attempt to theory specification by enriching the independent variables while retaining the prior dependent variable, the quality of ranking (here, e.g., instance recall and instance precision). Thus a major goal in such a theory is find out, which of the IR techniques yields the best ranking in the hands of a human searcher.

6.3 Theory of Information Access

Later interactive IR studies left the sole focus on the ranked list quality and aimed to establish searcher effectiveness in accessing information. We call theories in this range as theories of information access (Figure 13). Such theories seek to explain information access related to a range of

[3] See TREC Interactive track publications at http://trec.nist.gov/pubs.html

techniques, systems or approaches, i.e. all information seeking. In this case, the possible variables to be explained grow in number and type. They include satisfaction, access behavior (process), quality of information, and learning. Their variation may be explained by a combination of knowledge about the usual suspects: topic, process, collection, and system features. Further variables such as knowledge of information ecology, barriers, and objective information ecology may also be used.

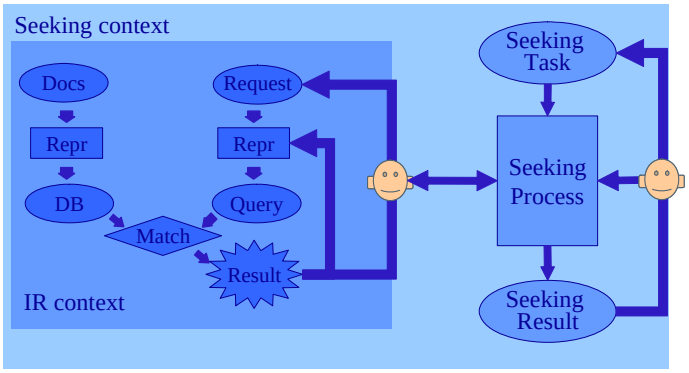

Fig. 13. Components of the theory of information access

Almost needless to say, such a comprehensive theory does not exist. Because there are many variables involved, the study designs are complex. Several studies have analyzed searcher's satisfaction and effectiveness, (dependent; e.g. number of documents saved, time used, satisfaction), the searcher, or the process (independent; time allocated, number of iterations) in addition to ranking (e.g. [2] [13] [23] [24]). The results have been mixed. These issues do not reduce the significance of pursuing such a theory, even if only piece-by-piece. Such a theory would be helpful in designing IR systems that fit the chosen information ecology.

6.4 Theory of Information Interaction

Our last context deals with the ultimate outcomes of information access: whether information access improves task performance or not (Figure 14). In theories of this sort the dependent variables are related to task performer's satisfaction, task process (e.g., duration, costs, quality), and task result (e.g., quality). The independent / controlled variables relate to task type, stage, and complexity; the variables from the theories above; performer's initial knowledge on task, see e.g. [15].

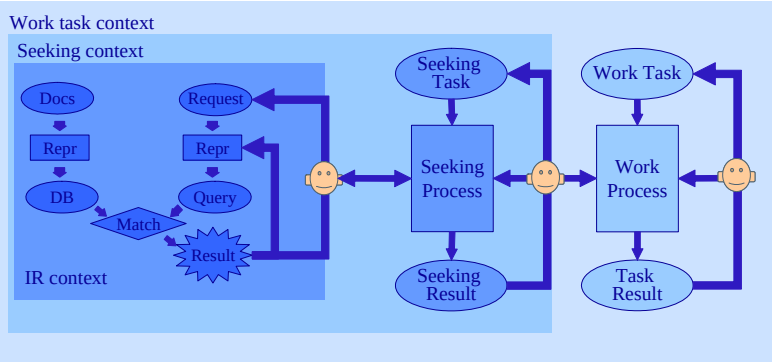

Fig. 14. Components of the theory of information interaction

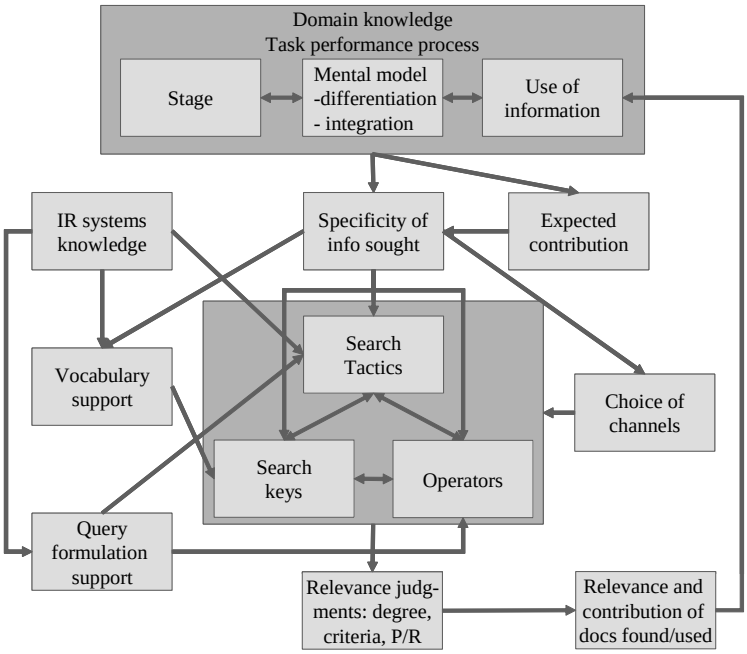

Fig. 15. Vakkari's theory of task-based information searching [25]

Vakkari [25] proposes a theory of task-based information access (Figure 15). It relates persons in task performance process and results with associated information access. A task proceeds in stages and the performer has a mental model of each task. The integration and differentiation of that model varies. Information is used within the process. The task performer has conceptions on information needs and searching, and expectations of contributions. There are system functionalities, which produce outputs, relevance and contributions of documents or their information.

I'm not suggesting this as THE THEORY OF INFORMATION RETRIEVAL. Rather, I'm inviting everyone to consider what it tries to accomplish. It relates task features, systems features, system uses, and outputs. Not just listing stakeholders in an overview manner; but detailing the effects. It is this type of theories that have a chance to explain, how information systems and access contributes to task performance. Let us develop them. There certainly lie difficulties and complexities ahead. I wonder how much progress we might have in developing such theories if equally many researchers studied them as there are researchers studying the theory of ranking.

Would this kind of an approach sacrifice THE SCIENCE OF IR, rigorous experiments as our hallmark? Many influential researchers in IR have warned us of opening this Pandora's box of user and task based approach as a great risk. This would release all evils and enemies of good science to surround us. We would not be able to argue rationally whose engine is the best.

7 Discussion and Conclusion

I think it is better to face the challenges in developing broader theories of information access and interaction, because they don't go away even if we cannot or don't want to see them. The situation can be called as: W Y D S I W Y D U = What you don't see is what you don't understand. If we close our eyes for broader frameworks, we may be like blind men making theories about elephants. Or, if we do not see the real life use of IR systems and other systems for task performance, we risk our possibilities of learned design of truly useful systems.

A group of prominent IR researchers produced the Meeting of the MINDS report: An Information Retrieval Research Agenda [9]. The report looked at recent developments in the field, challenges we are facing and proposed a research agenda. Here are some of the challenges:

- **Heterogeneous Data:** IR systems must seamlessly integrate and correlate information across a variety of media, sources, and formats.
- **Heterogeneous Context:** Search engines are context free. Need to understand the user, the domain and the larger task. Search is not the end goal.
- **Beyond the Ranked List:** People move from finding documents toward information interaction. Tools are needed for information transformation: clustering, linking, highlighting social networks, summarizing and arranging.
- **What Do People Really Do?** Little is known about how people use information retrieval tools. We are bound to research methodologies defined in the 1970s.
- **Evaluation:** Evaluation of new tools will require development of new metrics and methodologies.

To meet these challenges, at least three approaches are required:

- Development of evaluation methodologies - I believe that we would often be willing to do some novel type of research, but choose not to do that, because we are frustrated by not knowing how to evaluate the results under the Cranfield Paradigm, and knowing that we therefore cannot publish the result. The requirement for specific type of evaluation becomes here a straight-jacket for informative research.
- Descriptive empirical studies – lots of. If we don't know what is happening out there we cannot properly evaluate. If we want to be scientists in addition to being engineers, we need to understand the phenomena we are dealing with even if everything could not be immediately turned into tools and evaluated. Understanding feeds evaluation.
- Theories in broader scopes – there is neither a single theory nor a single type of theory for IR. We need to carefully think, what we want to understand and explain through our theories. In the long run we need theories that relate information organization and retrieval tools to information interaction in tasks (or leisure time). Otherwise we are paying lip-service to the goals of IR research. Theories help focus design efforts effectively.

It is important to evaluate the subsystems of information retrieval processes, including the search engines. By lifting one's eyes from result ranking quality alone, one may be able to put the subsystems and their contributions in relation with each other. Consequently, one may readily understand that much of the IR terrain is still unmapped, even from the systems viewpoint. Solving easy problems does not make great history. Solving the difficult ones matters. Task-based and user-oriented evaluation/research offer such problems. Solving them can potentially lead to significant progress in the field.

8 References

[1] Allan, J., Carterette, B., Lewis, J.: When will information retrieval be "good enough"? In: Proc. ACM SIGIR'05, pp. 433--440, Salvador. ACM (2005)

[2] Al-Maskari, A. & al.: The good and the bad system: Does the test collection predict users' effectiveness? In: Proc. of

ACM SIGIR'08, pp. 59--66, Singapore. ACM (2008)

[3] Armstrong, T.G. & al.: Improvements That Don't Add Up: Ad-Hoc Retrieval Results Since 1998. In: Cheung, D. & al (Eds.) Proceedings of the 18th ACM CIKM'09, pp. 601--610 Hong Kong. ACM (2009)

[4] Bates, M.J.: Where should the person stop and the information search interface start? Information Processing & Management 26(5): pp. 575--591 (1990).

[5] Belkin, N.J. & al.: Distributed expert-based information systems: An interdisciplinary approach. Information Processing & Management 23(5): pp. 395--409 (1987)

[6] Belkin, N.J., Croft, W.B.: Retrieval techniques. In: Williams, M.E. (Ed.) Annual Review of Information Science and Technology, vol. 22: pp. 37-61. Elsevier (1987)

[7] Belkin, N.J., Oddy, R.N., Brooks, H.M.: Ask for information retrieval: Part 1. Journal of Documentation 38(2): pp. 61--71 (1982)

[8] Borgman, C.L.: Why are online catalogs still hard to use? Journal of the American Society for Information Science 47(7): 493--503 (1996)

[9] Callan, J. & al.: Meeting of the MINDS: An information retrieval research agenda. SIGIR Forum 41(2): pp. 25--34 (2007)

[10] Englbart, D.C.: Augmenting Human Intellect: A conceptual Framework. Stanford Research Institute (1962)

[11] Fidel, R.: Moves in online searching. Online Review 9(1): pp- 61--74 (1985)

[12] Harman, D.: Information retrieval evaluation. Morgan & Claypool (2011)

[13] Huuskonen, S., Vakkari, P.: Students' search process and outcome in Medline in writing an essay for a class on evidence based medicine. Journal of Documentation 64,: pp. 287--303 (2008)

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

[15] Ingwersen, P., Järvelin, K.: The Turn: Integration of Information Seeking and Retrieval in Context. Springer (2005)

[16] Järvelin, K.: An Analysis of Two Approaches in Information Retrieval: From Frameworks to Study Designs. Journal of the American Society for Information Science and Technology 58(7): pp. 971--986 (2007)

[17] Järvelin, K.: Explaining User Performance in Information Retrieval: Challenges to IR evaluation. In: Azzopardi, L. & al. (eds.) Proceedings of the 2nd International Conference on the Theory of Information Retrieval, pp. 289--296, Cambridge. Springer, LNCS, vol. 5766 (2009)

[18] Kekäläinen, J. & Järvelin, K.: Evaluating information retrieval systems under the challenges of interaction and multi-dimensional dynamic relevance. In: Bruce, H. et al. (Eds.) Emerging Frameworks and Methods. Proceedings of the 4th International Conference on Conceptions of Library and Information Science (CoLIS 4), pp. 253--270, Seattle. Libraries Unlimited (2002)

[19] Kumpulainen, S., Järvelin, K.: Information Interaction in Molecular Medicine: Integrated Use of Multiple Channels. In: Belkin, N. & al. (eds.) Proceedings of the Information Interaction in Context Conference (IIiX 2010), pp. 95--104, New Brunswick. ACM (2010)

[20] Kumpulainen, S., Järvelin, K.: Barriers to Task-based Information Access in Molecular Medicine. Journal of the American Society for Information Science and Technology (JASIST) 62(x): to appear. (2011)

[21] Robertson, S.: On theoretical argument in information retrieval. SIGIR Forum 34(1): pp. 1--10 (2000)

[22] Saracevic, T.: User lost: Reflections on the past, future, and limits of information science. ACM SIGIR Forum 31(2) : pp.16--27 (1997)

[23] Smith, C.L., Kantor, P.B.: User Adaptation: Good Results from Poor Systems. In: Proc. ACM SIGIR'08, pp. 147--154, Singapore. ACM (2008)

[24] Turpin, A., Scholer, F.: User performance versus precision measures for simple search tasks. In: Proceedings of ACM SIGIR'06, pp. 11--18, Seattle. ACM (2006)

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

[26] Vakkari, P., Huuskonen, S.: Search effort degrades search output, but improves task outcome. Journal of the American Society for Information Science and Technology (JASIST) 62(x): to appear. (2011)