Some(what) Grand Challenges for Information Retrieval[1]

Nicholas J. Belkin
School of Communication, Information and
Library Studies
Rutgers University
New Brunswick, NJ, USA
nick@belkin.rutgers.edu

Abstract

Although we see the positive results of information retrieval research embodied throughout the Internet, on our computer desktops, and in many other aspects of daily life, at the same time we notice that people still have a wide variety of difficulties in finding information that is useful in resolving their problematic situations. This suggests that there still remain substantial challenges for research in IR. Already in 1988, on the occasion of receiving the ACM SIGIR Gerard Salton Award, Karen Spärck Jones suggested that substantial progress in information retrieval was likely only to come through addressing issues associated with users (actual or potential) of IR systems, rather than continuing IR research's almost exclusive focus on document representation and matching and ranking techniques. In recent years it appears that her message has begun to be heard, yet we still have relatively few substantive results that respond to it. In this paper, I identify and discuss a few challenges for IR research which fall within the scope of association with users, and which I believe, if properly addressed, are likely to lead to substantial increases in the usefulness, usability and pleasurability of information retrieval.

1 Section

As I looked over the Programme for ECIR 2008, I noted that one of the Tutorials, none of the Workshops, none of the Session titles, none of the Long Papers, none of the Short Papers, and none of the Posters has the word "user" or its cognates in the title. Although this particular fact may not be an entirely accurate reflection of the attention paid by the contributors to ECIR 2008 to user-oriented issues in IR (*personalization*, for instance, or *social media* are likely to be concepts indicative of such concern), it surely does tell us something of the general status of such issues within this research community. I don't mean to focus explicitly on ECIR in this respect (the Proceedings of the SIGIR conferences have precisely the same character), but to use the example to indicate what I perceive to be the general tenor of contemporary mainstream research in IR.

It's surprising to me that this is the case, as for quite some time various important representatives of the IR research community have suggested that there would be significant payoff to shifting emphasis in research from issues of representation and retrieval techniques, and developing and refining new models of IR, to research concerned with how to take account of the user in the IR system. I think, for instance, of Karen Spärck Jones's speech in acceptance of the ACM SIGIR Gerard Salton Award in 1988 (Spärck Jones, 1988), where she stated, with respect to her own very substantial research in relevance feedback that:

[1] This is a slightly revised version of a Keynote Lecture presented at the European Conference on Information Retrieval, Glasgow, Scotland, 31 March 2008.

> More importantly, I felt that the required next step in this line of work was to carry out real, rather than simulated, interactive searching, to investigate the behaviour of relevance weighting under the constraints imposed by real users, who might not be willing to look at enough documents to provide useful feedback information. (p. 18)

She went on to say (referring to IR research at least through the 1970s):

> But these concerns, though worthy, had unfortunate consequences. One was that, in spite of references to environmental parameters and so forth, it tested information systems in an abstract, reductionist way which was not only felt to be disagreeably arid but was judged to neglect not only important operational matters but, more importantly, much of the vital business of establishing the user's need. … The rather narrow view was however also a natural consequence of the desperate struggle to achieve experimental control which was a very proper concern and which remains a serious problem for IR research, and particularly the work on interactive searching to which I shall return later.
>
> As it is, it is impossible not to feel that continuing research on probabilistic weighting in the style in which it has been conducted, however good in itself in aims and conduct, is just bombinating in the void …
>
> The current interest, clearly stemming from the growth of computing power and the extension of computer use, is in integrated, personalisable information management systems. (p.18)

Some years later, in accepting the same award at the 1997 SIGIR meeting, Tefko Saracevic (Saracevic, 1997) stressed the significance of integrating research in information seeking behavior with research in IR system models and algorithms, saying:

> *… if we consider that unlike art IR is not there for its own sake, that is, 1R systems are researched and built to be used, then IR is far, far more than a branch of computer science, concerned primarily with issues of algorithms, computers, and computing*. (His emphasis) (p. 17)

And in discussing the goal of IR as helping people to find relevant information, and thus how to construe information and relevance, and what that means for IR:

> Broadest sense: Information that involves not only messages (first sense) that are cognitively processed (second sense), but also *a context - a situation, task, problem-at-hand, the social horizon, human motivations, intentions*…. (His emphasis)

> For information science in general, and IR in particular, we have to use the third, broadest interpretation of information, because users and use are involved - and they function within a context. That's what the field and activity is all about. That is why we need to consider IR in the broader context of information science. (p. 18)

Most recently, Ingwersen and Järvelin (2005), in their book, *The Turn*, have proposed a program for integrating studies of users, their tasks, goals, information problems and behaviors with research in IR techniques. It is still a bit early to evaluate the influence of their proposal (indeed one might say, manifesto) on IR research, but we can at least hope for a more immediate response than that given to the arguments of Spärck Jones and of Saracevic.

Certainly, I don't mean to suggest that these calls for a shift in the IR research paradigm have gone completely unanswered, as, for instance, research here at Glasgow University amply demonstrates, and as does the

appearance of the Information Interaction in conteXt (IIiX) conference series. Nevertheless, we can still see the dominance of the TREC (i.e. Cranfield) evaluation paradigm in most IR research, the inability of this paradigm to accommodate study of people in interaction with information systems (cf. the death of the TREC Interactive Track), and a dearth of research which integrates study of users' goals, tasks and behaviors with research on models and methods which respond to results of such studies and supports those goals, tasks and behaviors.

This situation is especially striking for several reasons. First, it is clearly the case that IR as practiced is inherently interactive; secondly, it is clearly the case that the new models and associated representation and ranking techniques lead to only incremental (if that) improvement in performance over previous models and techniques, which is generally not statistically significant (e.g. Spärck Jones, 2005); and thirdly, that such improvement, as determined in TREC-style evaluation, rarely, if ever, leads to improved performance by human searchers in interactive IR systems (e.g. Turpin & Hersh, 2001; Turpin & Scholer, 2006).

I know that I'm not telling this audience something that's radically new, either in an historical context, or in your understanding of the fundamental problems of IR. But, on the evidence, it seems to me that the field is still somewhat adrift in applying such understanding to what we accept as significant research problems, and appropriate ways in which to address such problems. Thus, what I'd like to do here is to propose what I consider to be a few outstanding such problems, or *challenges* for the field, and how they might be addressed.

2 Challenges

2.1 Information-related goals, tasks and intentions

There is substantial and convincing evidence that the goals that lead people to engage in information behavior, the tasks associated with those goals, and with their behaviors, and the intentions underlying the behaviors, substantially affect their judgments of usefulness of information objects, and the ways in which they interact (or would wish to interact) with information objects. The challenges here lie in three spheres.

First, in the ability to characterize and differentiate among information-related goals, tasks and intentions in some principled manners that go beyond straightforward listings, that will apply across a wide variety of contexts, and from which design principles for IR systems can be inferred. Although there have been studies which have characterized tasks on one or two dimensions, our fundamental understanding of these issues remains impoverished. And although there have been studies of the relationships between specification of tasks and intentions and information behaviors, these have yet to lead to explicit design principles for IR systems.

Second, we need to develop methods for inferring information-related goals, tasks and intentions from implicit sources of evidence, such as previous or concurrent behaviors. Without such techniques, any characterization of these factors, or even specification of how to respond to them in system design, is fruitless. At the moment, almost all research in this area is dependent upon explicit specification of goals, tasks and intentions, and there seem to be, as yet, no substantive proposals for how to perform such inference.

Third, going from characterization and identification of goals, tasks and intentions, to IR techniques which actually respond effectively to them, is a challenge that has been barely noted, much less addressed, to date. One reason is, of course, that we so far don't have the necessary characterizations, but another is that putting together the research expertise in the study of information behavior and in the development of IR techniques is a challenge in and of itself. In particular, the disciplinary structures in which these different types of research are carried out tend strongly to work against the necessary collaboration. Although such barriers may be

breaking down, there remain serious obstacles to the type of interdisciplinarity that seems to be required to meet this challenge, at least in the academic environment.

2.2 Understanding and supporting information behaviors other than specified search

People engage in a wide variety of types of interactions with information, of which specified searching, as represented by, e.g., normal Web search engines, and standard IR models, is only one among many, and perhaps not the most important. For instance, since we know, from a good number of different theoretical and empirical studies, that people have substantial difficulties in specifying what it is (that is, what information objects) that would help them to realize their goals, only considering specified search as the basis for IR models and techniques, is clearly inadequate, and inappropriate. Furthermore, a number of studies have demonstrated that people do engage in a variety of different information behaviors within even a single information seeking episode (e.g. Cool & Belkin, 2002; Olston & Chi, 2003).

Although there have been some attempts to support different information behaviors in an integrated manner, such as support for searching and browsing, there is still little known about the nature, variety and relationships of different information behaviors, and the situations that lead people to engage in any one behavior, or sequence of behaviors. Without this basic knowledge, there is little hope for the development of systems which can proactively support multiple behaviors. Yet, we need to have experimental conditions in which system functionalities can be evaluated with respect to the ways in which they support different behaviors, and sequences of behaviors, and in which the independent variables are situational characteristics which can be manipulated and whose effects can thus be investigated. This clearly calls for integration of the information seeking and IR techniques research paradigms.

2.3 Characterizing context

At the 2006 IIiX conference, there was a panel session entitled: "What is not context?" The intent of the panel was to see if there were some way to identify some subset, of all of the possible factors leading to and surrounding an information seeking situation, which are necessary and sufficient in considering what an IR system should do in order to be contextually responsive. The panel's (and audience's) perhaps unfortunate conclusion was that *everything* is context, and that, at least at the moment, there appears to be no principled way in which to narrow to some subset of such factors. In summarizing the discussion, it seemed that the best that can be done is to try to identify aspects of context, knowledge of which seem likely to have the largest effects on improving support for information behaviors, and to devise and evaluate techniques for taking account of each aspect.

The challenges here are obvious, and multiple. We need to have theories and data which lead not only to identification of at least classes of contextual features, but which also tell us in what ways these features do in fact affect information behaviors. We need to have ways to identify appropriate contextual features without explicitly eliciting them. We need to develop IR system techniques which can actually take account of knowledge of these features, individually and in combination, and we need to develop experimental methods and evaluation measures which will allow comparison of the effect of contextually-based interventions. Once again, it seems that integration of the two research paradigms is called for, even to get started on this overall challenge.

2.4 Taking account of affect

To date, mainstream IR research, most research in interactive IR, and indeed most information behavior research, have been concerned primarily with efficiency and effectiveness (from a cognitive perspective) of the IR system, or of the performance of the user in the IR system. Despite Carol Kuhlthau's research demonstrating the significance of affect in the Information Seeking Process already some seventeen years ago (Kuhlthau, 1991), there has been almost no serious research effort in understanding the role of affect in the

information seeking situation in general and the IR situation in particular, nor in IR system design. Admittedly, concern with so-called affective computing is relatively new in general, so we cannot say that IR is alone in its failure even to acknowledge the significance of this aspect of the user's experience in the IR system. A recent collection of papers is an important indication that this situation is beginning to change, at least in the arena of information behavior research (Nahl & Bilal, eds, 2007), and Arapakis & Jose (2008) is a significant first step in this direction in IR research.

Intuitively, we recognize that the emotions which people experience during interaction with information, whether in an IR system or not, can substantially affect how such interaction proceeds, and how subsequent interactions might take place. Yet we know very little even about *what* affective experiences people undergo during the course of information interaction, much less what might cause such experiences, and with what effects, nor what emotional experiences might be desirable or "useful" during an interaction.

Making IR pleasurable is surely as significant a goal as making it cognitively or situationally effective, and indeed might influence attainment of these types of effectiveness. But we can surely reason, based on the little research in information behavior on this issue, that making the entire IR experience pleasurable may not imply only happy emotions throughout the course of any one information interaction episode.

Our challenges here are first even to recognize the significance of the issue; to understand what emotions do occur during information interaction; to determine ways in which to reliably recognize affective states during interactions; to learn the causes and effects of these affective states; and finally to design IR techniques which can appropriately reduce the occurrence of undesirable affective responses, to encourage the incidence of desirable affective states, and to take advantage of knowledge of affective states in order to enhance the entire information interaction experience.

2.5 Personalization

Personalization of IR is clearly an active and significant aspect of IR and information seeking research already. However, the typical way in which personalization is construed seems actually rather narrow. Most research in this area is concerned with identifying an aspect of current or past behaviors, such as dwell time on a page, click-through, information objects which have been previously been viewed or saved or used, and using that information as a form of relevance feedback, to either modify in some way an initial query, or to rank the results of a query in a way tailored to that knowledge (cf. Kelly & Teevan, 2003; Kelly, 2005).

This type of approach is narrow because it typically uses only one type of evidence in isolation (although recent research is beginning to look at the interactions of such evidence); because the type of evidence considered is typically behavior with respect to information objects; and because personalization is limited to a single aspect of the information seeking situation, the prediction of relevance of information objects. There is still rather preliminary research which indicates that there are strong interaction effects amongst different types of evidence that affect their interpretation as indicators of relevance (e.g. Kelly & Belkin, 2004; White & Kelly, 2006; Teevan, Dumais & Liebling, 2008). It is also the case that there is good reason to suppose that factors associated with goals, tasks, the individual's knowledge, and a variety of other contextual features can also affect the interpretation of behaviorally-based evidence. And, it is clear that prediction of relevance of information objects specific to a person's current situation is only one aspect of what one might consider a truly personalized information interaction.

So, the challenge with respect to personalization is first to consider the dimensions along which personalization could and should take place; then to identify the factors or values in each dimension that would inform personalization on that dimension; then to investigate how the different factors or types of evidence

interact with one another; and finally, to devise techniques which take account of these results in order to lead to a really personalized experience.

2.6 Integration of IR in the task environment

Engaging in information interaction in an IR system is almost always a consequence of a person's wanting to achieve some other goal, or accomplish some other task. In this sense, such engagement is inevitably a distraction from that other task. So, we might say that an ultimate goal of, and challenge for IR research is to arrange things such that a person never has to engage in interaction with a separate IR system at all (although I'm quite willing to agree that there are certainly circumstances in which such engagement might indeed be desirable).

For quite some time now, various researchers have recognized the significance of this issue, and investigated methods for integrating IR within various task environments, so that people do not have to leave their tasks in order to have their information problems resolved (e.g. Budzik & Hammond, 2000). However, to my knowledge none of such efforts have gone beyond the experimental stage, and few if any have been evaluated in any serious way. What would it take to get beyond such initial attempts to truly effective integration of IR in the task environment?

This is surely a serious challenge, for it depends upon deep understanding of the nature of the variety of tasks which might lead to information problems, and of the nature of the information problems which people might encounter in such tasks. Beyond such understanding, achieving this goal clearly requires collaboration with those who construct the tools which support the tasks themselves. All this may mean that there may be no general solution to the goal of integration in the task environment, but rather that we will need to build such integration into each support application, as a matter of course. Of course, this is in a sense a general solution, but one which cannot be accomplished by the IR community on its own, but only in collaboration not only with the information behavior community, but with the application communities. Perhaps there may be general principles which can be applied across task contexts, and even IR modules which might be used across applications, but we will not know whether this is indeed the case unless we substantially increase our research efforts in integration of IR into a number of specific, real task environments, with the concomitant application of the methods of information behavior and IR technique paradigms.

2.7 Evaluation paradigms for interactive IR

Speaking of paradigms, as Spärck Jones pointed out twenty years ago, well before TREC, the TREC evaluation paradigm is quite unsuitable for evaluation of interactive IR systems. As she also pointed out, the kinds of methods that would seem to be most suitable for this purpose have severe limitations or constraints, which appear to be extremely difficult to overcome. These have mostly, although not exclusively, to do with the inherent non-replicability of interactive IR.

This issue has of course not gone unnoticed (cf. Robertson & Hancock-Beaulieu, 1992), and there have been attempts at developing experimental methods and evaluation measures which are more suitable to the interactive IR situation. The TREC and CLEF and INEX Interactive Tracks are examples of such attempts, as is the evaluation model proposed by Borland (2003), among others. Indeed, there are contributions to ECIR 2008 which attempt to address various aspects of this general problem (e.g. Järvelin, et al., 2008). But it seems that up to now, we have not been successful at resolving the contradictions between the necessity for realism, and the desire for comparability and generalization.

I would like to suggest that there are two possibly interesting approaches to addressing this general issue. One is dependent on the enormous masses of data on interactive searching that are collected and analyzed by Web

search engines. We have all seen examples of the kind of qualitative differences in analysis and understanding of search that are made possible by the sheer quantitative differences that are now available. And there are clearly ways in which to integrate the analysis of such logs with selective investigations of search behaviors and intentions that could give even more nuanced results. But there are also clearly problems with the general research community's being able to make us of these data, as, for instance, the AOL experience reminds us.

A possible alternative might be the establishment of some minimal standards for collection of data in interactive IR experiments, such that the records of the interactions might be cumulated, constituting eventually a kind of "test collection" of IR interactions, on which different analyses and different experiments might be conducted. Such a collection, while addressing the problems of generalizability and cost of experimentation, would still not necessarily address the issue of validity, but it might be an interesting and useful start toward that goal.

2.8 (In)Formal models of interactive IR

There are a number of papers at this conference, and similar conferences, such as SIGIR, that are concerned with the introduction of new, more or less formal models of IR (e.g. language modeling), refinement of such models (e.g. smoothing), implementation of the models in specific techniques, and evaluation of the effectiveness of the techniques, and therefore, it is claimed, of the validity (or at least usefulness) of the models themselves. This is all interesting and useful work, which nevertheless appears to me to suffer from much the same problems identified by Spärck Jones these twenty years ago. That is (and this may well be a contentious claim), none of these models has any place for the user, nor for interaction. All of them are primarily, indeed usually *only* concerned with issues of representations of information objects and given, static queries, and of matching and ranking techniques. As the exception to prove the rule, Norbert Fuhr (2008) has recently published a paper outlining what I believe to be the first attempt at developing a truly formal model of interactive IR.

3 Conclusion

I'd like to suggest that a challenge underlying all of the challenges that I've outlined here is the challenge to address seriously the inherently and inevitably interactive nature of IR, by moving beyond the limited, inherently non-interactive models of IR that we have been concerned with, to the development of models of IR which incorporate the user as an active participant in the IR system, and which treat the person's interaction with information as a central process of IR. This will require the joint efforts of contributors from a variety of research traditions, and may mean that we will have to give up the idea of strictly formal models of IR, in favor of realistic and useful models of IR. This, in my opinion, may not be a bad trade off.

4 References

Arapakis, I. & Jose, J. (2008) Affective Feedback: An investigation of the role of emotions during an information seeking process. In *SIGIR 2008. Proceedings of the 31st Annual ACM SIGIR International Conference on Research and Development in Information Retrieval* (in press). New York: ACM.

Borlund, P. (2003). The IIR Evaluation Model: a Framework for Evaluation of Interactive Information Retrieval Systems. In: *Information Research*, vol. 8, no. 3, paper no. 152. [Available at: http://informationr.net/ir/8-3/paper152.html]

Budzik, J. and Hammond, K. J. (2000) User Interactions with Everyday Applications as Context for Just-in-Time Information Access. In *IUI 2000, ACM Conference on Intelligent User Interfaces* (pp.44-51). New York: ACM.

Cool, C. & Belkin, N. J. (2002). A classification of interactions with information. In Proceedings of the Fourth International Conference on Conceptions of Library and Information Science (pp. 1-15). Greenwood Village, CO: Libraries Unlimited.

Fuhr, N. (2008) A probability ranking principle for interactive information retrieval. *Information Retrieval*, v. 11: 251-265.

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

Järvelin, K., Price, S.L., Delcambre, L.M.L. & Nielsen. M.L. Discounted Cumulated Gain Based Evaluation of Multiple-Query IR Sessions. In *ECIR 2008, Proceedings of the 2008 European Conference on Information Retrieval* (pp. 4-15). Berlin: Springer Verlag.

Kelly, D. (2005). Implicit feedback: Using behavior to infer relevance. In A. Spink and C. Cole (Eds.) *New Directions in Cognitive Information Retrieval* (pp. 169-186). Berlin: Springer Verlag.

Kelly, D. & Belkin, N.J. (2004). Display time as implicit feedback: Understanding task effects. In *SIGIR 2004, Proceedings of the 27th Annual ACM International Conference on Research and Development in Information Retrieval* (pp. 377-384). New York: ACM.

Kelly, D. & Teevan, J. (2003). Implicit feedback for inferring user preference: A bibliography. *SIGIR Forum*, 37(2), 18-28.

Kuhlthau, C. C. (1991). Inside the search process: information seeking from the user's perspective. *Journal of the American Society for Information Science, 42*, 361-371.

Nahl, D. & Bilal, D. eds (2007) Information *and emotion: The Emergent Affective Paradigm in Information Behavior Research and Theory*. Medford, NJ: Information Today for ASIST.

Olston, C. & Chi, Ed H. (2003). ScentTrails: Integrating browsing and searching on the web. *ACM Tarnsactions on Computer-Human Interaction, 10*(3), 177-197.

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

Saracevic, T (1997) Users lost: reflections of the past, future and limits of information science. *SIGIR Forum*, 31, 2: 16-27.

Spärck Jones, K. (1988) A look back and a look forward. In: *SIGIR '88. Proceedings of the 11th Annual ACM SIGIR International Conference on Research and Development in Information Retrieval* (pp. 13-29). New York: ACM.

Spärck Jones, K. (2005). Meta-reflections on TREC. In E.M. Voorhees & D.K. Harman (Eds.) *TREC: Experiment and Evaluation in Information Retrieval* (pp. 421-448). Cambridge, MA: MIT Press.

Teevan, J., Dumais, S.T. & Liebling, D. J. (2008) To personalize or not to personalize. In *SIGIR 2008. Proceedings of the 31st Annual ACM SIGIR International Conference on Research and Development in Information Retrieval* (in press). New York: ACM.

Turpin, A. H. & Hersh, W. (2001) Why batch and user evaluations do not give the same results. In *SIGIR 2001, Proceedings of the 24th Annual ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 225-231). New York: ACM.

Turpin, A. & Scholer, F. (2006) User performance versus precision measures for simple search tasks. In *SIGIR 2006, Proceedings of the 29th Annual ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 11-18). New York: ACM.

White, R. W. & Kelly, D. (2006). A study on the effects of personalization and task information on implicit feedback performance. In *CIKM '06, Conference on Information and Knowledge Management* (pp.). New York: ACM.