On the evaluation of interactive information retrieval systems

Citation to Publisher Version: No citation available.

Citation to *this* Version: Belkin, Nicholas J.. *On the evaluation of interactive information retrieval systems*, 2010. Retrieved from doi:10.7282/T3SF2TJK.

On the Evaluation of Interactive Information Retrieval Systems

Nicholas J. Belkin

Department of Library and Information Science, School of Communication & Information,
Rutgers University, 4 Huntington Street, New Brunswick, NJ 08901, USA
belkin@rutgers.edu

Abstract. This paper briefly discusses the history of the standard information retrieval evaluation criteria, measures and methods, and why they are unsuitable for the evaluation of interactive information retrieval. A new framework for evaluation of interactive information retrieval is proposed, based on the criterion of usefulness.

Keywords: Interactive information retrieval, information retrieval evaluation.

1 Introduction

It is both a great honor, and a great pleasure for me to contribute to this celebration of the career of my long-time friend and colleague, Peter Ingwersen. Furthermore, it turns out to be, at least in one respect, a relatively easy task, in that Peter has made significant contributions in so many areas of information science, that finding a topic both relevant to his interests, and to my current research concerns, is not a great problem. Of more moment, of course, is to achieve his level of insight.

Among Peter's continuing concerns has been the evaluation of interactive information retrieval systems (e.g. [1] [2]), and it is this particular issue that I wish to address in this paper. For well on 20 years now (see, e.g. [3]), it has been quite clear that the standard Cranfield/TREC model of information retrieval (IR) system evaluation is very badly suited to the evaluation of interactive IR systems. Since IR is an inherently interactive activity, from a theoretical point of view (e.g, [4]), and has been from a practical point of view since the 1970s, it is a severe problem that almost all criteria, measures and methods used in formal IR system evaluation continue to be those which have been designed to test non-interactive IR.

In this paper, I discuss just why the standard IR evaluation criteria, measures and methods are not suited, in the general case, to the evaluation of interactive IR (IIR), suggest that the criterion of *relevance*, long held to be the central concept of IR, if not of information science itself (cf. [5]), is inappropriate (again, in the general case), and propose that considering the *usefulness* of an IIR episode, and of its components, with respect to its contribution to the accomplishment of the task that led to the episode, can lead to both realistic and informative evaluation of IIR systems.

2 Why have IR systems been evaluated as they have been?

There is a history to the evaluation of IR systems, and I believe that it is rooted in the practices of documentation, and especially of science librarianship. Bradford's discovery of bibliographic regularities arose through his analysis of the work that he did as a science librarian [6]. That work was the compilation of subject bibliographies, primarily on request of a scientist or a group of scientists. The goal of such bibliographies was to identify all of the documents pertaining to the subject, and to not include in the bibliography any documents which did not pertain to the subject. It is not difficult to see how Cyril Cleverdon, himself a science librarian (and others, of course), could accept these as goals for an IR system, understanding the phrase "pertaining to the subject" as meaning (eventually) "relevant to the inquirer's query", making relevance of a document the basic criterion of evaluation, and therefore leading to the measures of recall and precision, emulating the "all and only" of the subject bibliography.

The very first evaluations of IR systems, as at Cranfield [7] and Western Reserve [8], and their critics (e.g. Swanson, [9]), clearly recognized that there were some inherent problems with this general analogy, and with the concept of relevance, mostly having to do with the inherent subjectivity of relevance judgments. The response to these problems by the IR research community was to attempt to remove the person from the equation, thereby eliminating subjectivity. Both Cleverdon and his regular adversary, Jason Farradane [10] accepted that this was the only manner in which "scientific" evaluation of IR systems could be conducted.

Salton's SMART project recognized another difficulty with the standard model; that is, that a person's initial expression of an "information need" in some query was quite unlikely to be the best possible such expression. In Rocchio's [11] interpretation of this fact, the problem was seen as finding the "ideal" query, and the answer was for the IR system to interpret the searcher's evaluations of document relevance (or not) as evidence for query modification. Thus, there was implied in this formulation some idea of the searcher *interacting* with the IR system, but in a strangely passive mode. More substantive interaction, involving the searcher as an active participant, and also one whose information need, as represented by a query, might change through the course of an interaction, was explicitly not considered. Thus, the evaluation model, even in this partially interactive mode, remained the evaluation of the results of one specific query, with the same "all and only" measures.

3 Why shouldn't IR systems be evaluated as they have been?

The reasons which lead people to engage in information seeking, and therefore in interaction with information retrieval systems, seem only rarely to be equivalent to the goal of the subject bibliography (cf. [12] [13] [14] [15]). Indeed, a more apt example from the same era as Bradford's, might rather be the exploration of a library in order to discover relationships among ideas which one had not thought of before, such as interacting in the library of the Warburg Institute [16]; another might be to learn about a new domain of interest, through exploration of its canonical texts; yet another might be the desire to find one document which answers a specific question; a fourth could well be to obtain advice about possible courses of action in a given situation. It would be simple to continue this list for quite some time, if not quite endlessly. An alternative is to consider the possible circumstances underlying the *problematic situation*, as initially described in Schutz & Luckmann [17]), and applied in various ways to the contexts of information science and IR by, e.g., Belkin, Seeger & Wersig [18] Wersig [19]. Schutz & Luckmann quite plainly outline at least the knowledge-oriented reasons that might lead people to engage in information seeking; none of them, however, seems to lead to that which underlies the standard IR evaluation methods and measures. Even their quite extended and explicit discussion of relevance is of a concept quite different from that normally used in IR. Indeed, when considering the range of reasons that might lead people to engage with IR systems, we find that the situations in which finding all of the documents relevant to a query (or its underlying information "need") constitute a rather small minority, which suggests that a more general evaluation model, encompassing the range of reasons or goals of information seeking might be more appropriate.

It is also the case that many, if not most information seeking interactions take place not as isolated, single queries, but rather as information seeking episodes, during which various activities, including, but definitely not limited to the posing of different queries, take place (cf. Belkin, 1996 [20]: Fuhr, 2009 [21]). It thus makes sense to consider an evaluation paradigm which undertakes the evaluation of the search episode as a whole. But the relevance criterion and the "all and only" measures are suited (indeed designed) to evaluate the success of a single query, and it seems at the very least exceedingly difficult to adapt them to the evaluation of an entire search episode. The struggles, and eventual failure of the TREC Interactive Track Dumais and Belkin 2005 [22] in its attempt to evaluate IIR within the strictures of the standard evaluation paradigm give testimony to aspects of this problem. Järvelin, et al., 2008 [23] is an example, perhaps the only extant example, of an attempt at directly using relevance as the criterion for evaluation of an entire search episode, albeit with a quite different measure than recall or precision. The difficulties that they faced, and the problems that arose in the test of their measure and methods, illustrate the extreme difficulty of using relevance for this purpose. More often, when considering the evaluation of IIR, relevance and its companion measures have just been discarded, or, as in the TREC Interactive Track, supplemented by a variety of alternative measures. Su [24] suggested a measure which could, in principle, be applied to the entire search episode, "value of search results as a whole', which in fact does away completely with ideas of recall and precision, and perhaps even relevance, at least as commonly understood. Similarly, "satisfaction", measured according to multiple criteria, including satisfaction with the search episode (often operationalized as the interaction with a library and a librarian) has long been suggested (and used) as a more holistic criterion than just relevance for evaluation of IIR (e.g. Tagliacozzo [25]).

Furthermore, the nature of IIR is such that the information seeker's state of knowledge is quite likely to change during the course of the information seeking episode [14], leading to new ideas of what might be useful, as could even the person's understanding of the problem or task that led to information seeking [18]. As Bates [12] and Oddy [26] have proposed, just seeing some new text during the course of information seeking could lead to quite new ideas about what other texts it would be nice to encounter. But the only kind of interaction that the normal IR evaluation paradigm readily allows, relevance feedback leading to an ideal query, takes no account of these sorts of changes.

Thus, the standard IR evaluation paradigm fails to respond to the fundamental nature of IIR, in terms of the kinds of goals for information seeking that it presupposes, in terms of its inability to evaluate entire information seeking episodes, and in terms of its inability to account for the changes in the searcher that are inherent in interactive information seeking.

4 Usefulness as the criterion for evaluation of interactive information retrieval

Assume that the ultimate goal of IR is to support people in the resolution of their problematic situations [18] [20]. An operationalization of this goal that has been accepted by the IR community is the provision of texts relevant to a query. But quite different operationalizations can be, and have been imagined. Cooper [27], for instance, suggested that the *utility* of a search result is a more realistic criterion. My colleagues and I at Rutgers have questioned relevance as an appropriate criterion for evaluation of IIR, and suggested elsewhere that *usefulness* could be a much more realistic criterion [28] [29] [30]. Here, I draw on that work, sketching an outline of the argument in favor of usefulness, with some discussion of how it could be applied.

We begin by considering the issue of how to evaluate an IIR system in terms of the goal that we have assumed. The question that immediately arises is: how to relate what the system does (or doesn't do) to the resolution of the problematic situation. The issue here is how to know to what extent the problematic situation has been resolved; already in 1974, John Martyn [31] pointed out that our concern should be with the *use* of the information gained through interaction with the information system, yet we still lack methods, or a sound framework for directly understanding this relationship. One possibility for addressing this problem is to specify, quite concretely, the *task* which the searcher intends to accomplish, and then to measure to what extent, or how well that task has actually been accomplished, after the information retrieval interaction. To some extent, the method proposed by Borlund and Ingwersen [1] attempts to address this issue. The major difficulty remains the ability to establish a direct connection between what the system did, and what effect that had on the task outcome. Jean Tague's [32] proposal of a measure of *informativeness* was an early step in this direction, which has unfortunately not been followed up in subsequent research.

Our proposal for addressing this problem is to consider the *usefulness* of the IR interaction with respect to the motivating task at three distinct levels:

1. The usefulness of the entire interaction with respect to the motivating task;
2. The usefulness of each step in the information seeking episode with respect to accomplishing the goal of the interaction, and with respect to its contribution to accomplishment of the motivating task;

3. The usefulness of system support with respect to the goal of each individual step in the interaction.

Our contention is that, by decomposing the tasks/goals of an information seeking episode in this way, it will be possible to relate system support behaviors associated with each individual step during the course of the information seeking episode with the extent to which the motivating task has been resolved, combining both summative (motivating task) and analytic (individual step goals) evaluation methods.

The method, in the abstract, is as follows. First, the motivating task is elicited (in the case of participants searching for their own purposes) or controlled (as proposed in [1]), as are criteria and measures for evaluating the extent to which the task will be or has been accomplished, respectively. The goal of the information seeking episode itself is treated in the same manner. Then, the searcher engages in the IIR system, and the task (in the case of controlled searching) completed. All activities during the information seeking episode are logged/recorded.[1] At this point, task accomplishment is evaluated, and searcher evaluation of the usefulness of the information seeking interaction with respect to task accomplishment is elicited, as is the goal of the information seeking episode itself. Then, each step in the information seeking episode is examined, sequentially, eliciting from the searcher the goal of each step, in and of itself, and with respect to the accomplishment of the episode's information seeking goal, and the extent to which the goal of the specific step was achieved, and the usefulness of that step toward the accomplishment of the information seeking goal.

This procedure allows not only the establishment of the relationship of each support technique (associated with the individual steps) with the outcome of the searching process, and task accomplishment, but also can evaluate the sequencing of the steps, as a process leading to information seeking goal and task accomplishment. We have not considered in this description a number of factors that would need to be controlled or taken account of, in order to interpret the data appropriately. These would include, *inter alia*, characteristics of the searcher such as searching, topic and domain knowledge, cognitive abilities, and other individual differences. But we already have examples of how this could be done in a variety of IIR experiments.

Clearly, the method as outlined above is likely to be too cumbersome to be enacted in whole in a realistic (i.e. relatively large) evaluation exercise. But, one can imagine how various aspects of the evaluation could be accomplished without the great involvement of the searcher that is described. For instance, using the method of [1], suitably enhanced, can eliminate searcher involvement in the first step. Examining the search log to see what uses have been made of each step in subsequent steps could substantially reduce searcher involvement in evaluation of usefulness of each step toward the information seeking goal. Inferring individual step goals from the specific behaviors within each step, and applying appropriate evaluation measures, could again reduce searcher involvement. And, examining the sequence of steps for "aberrant" sequences (e.g. repetitions, backtracking) could inform the identification of an "ideal" sequence, and an evaluation of the system's support for helping the

[1] In the case of uncontrolled searching, at the end of the search, both motivating task and information seeking goal are again elicited, in order to confirm that they did not change; if they did change, we engage in the elicitation and measurement activity with respect to these, and consider when and why the changed in subsequent elicitation.

searcher to engage in that sequence. Of course, being able to do these sorts of abstractions will require substantial preliminary research using the full, searcher intensive method, but this should not deter us from moving toward the goal of truly good evaluation of IIR.

In summary, the criterion of usefulness, properly construed, can not only incorporate previous criteria, such as relevance, as special cases appropriate for evaluating specific steps within an information seeking episode, but also offers the opportunity to evaluate the effectiveness of an IIR system in such a way as to relate the support characteristics of that system to the success of the information seeking episode as a whole, in supporting the resolution of the searcher's problematic situation, and the accomplishment of the task that led the searcher to engage in information seeking behavior.

Acknowledgements

This research was supported by IMLS Grant LG-06-07-0105-07. Much of the content of this paper is due to my colleagues in the PoODLE Project[2] at Rutgers University, although I take sole responsibility for opinions expressed herein.

5. References

1. Borlund, P., Ingwersen, P.: The Development of a Method for the Evaluation of Interactive Information Retrieval Systems. Journal of Documentation. 53, 225-250 (1997)
2. Borlund, P., Ingwersen, P. Measures of relative relevance and ranked half-life: performance indicators for interactive IR. In: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, pp. 324-331. ACM Press, New York (1998)
3. Roberson, S.E., Hancock-Beaulieu, M.M.: On the Evaluation of IR Systems. Information Processing and Management. 28, 457-466 (1992)
4. Belkin, N.J.: Information Retrieval as Interaction with Information. In: Information Retrieval '93: von der Modellierung zur Anwendung, pp. 55-66. Konstanz, DE, Universitäts Verlag Konstanz (1993)
5. Saracevic, T.: Relevance: A Review of the Literature and a Framework for Thinking on the Notion in Information Science. Journal of the American Society for Information Science, 26, 321-343 (1975)
6. Bradford, S.C.: Documentation. C. Lockwood, London (1948)
7. Cleverdon, C.: The Cranfield Tests on Index Language Devices. Aslib Proceedings. 19, 173-194 (1967)
8. Saracevic, T.: An Inquiry into Testing of Information Retrieval Systems, Part i: Objectives, Methodology, Design and Control. Final technical report, Grant Phs Fr-00118 (1968).
9. Swanson, D. R.: Some Unexplained Aspects of the Cranfield Tests of Indexing Performance Factors. Library Quarterly. 41, 221-228 (1971)
10. Farradane, J.: The Nature of Information. Journal of Information Science, 1, 13-17 (1979).
11. Rocchio, J.J.: Relevance Feedback in Information Retrieval. In: Salton, G. (ed.) The SMART Retrieval System: Experiments in Automatic Document Processing, pp. 313-323. Prentice-Hall, Englewood Cliffs, NJ (1971)

[2] http://comminfo.rutger.edu/imls/poodle

12. Bates, M.J.: The Design of Browsing and Berrypicking Techniques for the Online Search Interface. Online Review, 13, 407-423 (1989)
13. Belkin, N.J.: Anomalous States of Knowledge as a Basis for Information Retrieval. Canadian Journal of Information Science. 5, 133-143 (1980)
14. Belkin, N.J., Oddy, R.N., Brooks, H.M.: ASK for Information Retrieval. Part I:. Background and Theory. Part II: Results of a Design Study. Journal of Documentation, 38, 61-71, 145-164 (1982)
15. Marchionini, G.: Exploratory Search: From Finding to Understanding. Communications of the ACM, 49 4, 41-46 (2006)
16. Gombrich, E.H.: Abby Warburg: An Intellectual Biography, 2nd edition. University of Chicago Press, Chicago (1986)
17. Schutz, A., Luckmann, T.: The Structures of the Life World. Northwestern University Press, Evanston, IL (1973)
18. Belkin, N.J., Seeger, T., Wersig, G.: Distributed Expert Problem Treatment as a Model for Information System Analysis and Design. Journal of Information Science, 5, 153-167 (1983)
19. Wersig, G.: Information – Kommunikation – Dokumentation. Pullach bei München, Verlag Dokumentation (1971)
20. Belkin, N.J.: Intelligent Information Retrieval: Whose Intelligence? In: ISI '96. Proceedings of the Fifth International Symposium on Information Science, pp. 25-31. Konstanz, DE, Universitäts Verlag Konstanz (1996)
21. Fuhr, N.: A Probability Ranking Principle for Interactive Information Retrieval. Information Retrieval. 11, 251-265 (2008).
22. Dumais, S.X., Belkin, N.J.: The TREC Interactive Tracks: Putting the User into Search. In: Voorhees, E.M., Harman, D.E. (eds.) TREC, Experiment and Evaluation in Information Systems, pp. 123-152. Cambridge, MA, MIT Press (2005)
23. Järvelin, K., Price, S.L., Delcambre, L.M.L., Lykke Nielsen, M.: Discounted Cumulated Gain Based Evaluation of Multiple-Query IR Sessions. In: Macdonald, C., et al. (eds.) ECIR 2008, LNCS 4956, pp. 4–15. Springer-Verlag, Heidelberg Berlin (2008)
24. Su, L.: Evaluation Measures for Interactive Information Retrieval. Information Processing and Management, 34, 557-579 (1998)
25. Tagliacozzo, R.: Estimating the Satisfaction of Information Users. Bulletin of the Medical Library Association, 65, 243-249 (1977)
26. Oddy, R.N.: Information Retrieval through Man-Machine Dialogue. Journal of Documentation, 33, 1-14 (1977)
27. Cooper, W.S.: On Selecting a Measure of Retrieval Effectiveness. Journal of the American Society for Information Science, 24, 87-100
28. Belkin, N.J., Bierig, R., Cole, M.: Is Relevance the Right Criterion for Evaluating Interactive Information Retrieval. In: Proceedings of the ACM SIGIR 2008 Workshop on Beyond Binary Relevance: Preferences, Diversity, and Set-Level Judgments. http://research.microsoft.com/~pauben/bbr-workshop (2008)
29. Belkin, N.J., Cole, M., Liu, J.: A Model for Evaluation of Interactive Information Retrieval. In: Proceedings of the ACM SIGIR 2009 Workshop on Understanding the User. http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-512/ (2009)
30. Cole, M., Liu, J., Belkin, N.J., Bierig, R., Gwizdka, J., Liu, C., Zhang, J., Zhang, X.: Usefulness as the Criterion for Evaluation of Interactive Information Retrieval. In: Proceedings of the third Workshop on Human-Computer Interaction and Information Retrieval. http://cuaslis.org/hcir2009/ (2009)
31. Martyn, J.: Information Needs and Uses. In: Cuadra, C. , Luke, A.W. (eds.) Annual Review of Information Science and Technology. 9, 3-24 (1974)
32. Tague-Sutcliffe, J.: Measuring the Informativeness of a Retrieval Process: In: Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 23-36. ACM Press, New York (1992)