Preprint: Järvelin, K. (2009). Explaining User Performance in Information Retrieval: Challenges to IR evaluation. In: L. Azzopardi & al. (Eds.), Proceedings of the 2nd International Conference on the Theory of Information Retrieval. 2009, Heidelberg: Springer, Lecture Notes in Computer Science vol. 5766, pp. 289-296.

Explaining User Performance in Information Retrieval: Challenges to IR evaluation

Kalervo Järvelin

University of Tampere, Finland
kalervo.jarvelin@uta.fi

Abstract. The paper makes three points of significance for IR research: (1) The Cranfield paradigm of IR evaluation seems to lose power when one looks at human instead of system performance. (2) Searchers using IR systems in real-life use rather short queries, which individually often have poor performance. However, when used in sessions, they may be surprisingly effective. The searcher's strategies have not been sufficiently described and cannot therefore be properly understood, supported nor evaluated. (3) Searchers in real-life seek to optimize the entire information access process, not just result quality. Evaluation of output alone is insufficient to explain searcher behavior.

1 Introduction

The dominant view on IR theory boils down to formal models of information retrieval (IR). These models are abstract specifications for the search engines to work – quite different from empirical theories, which one confirms or refutes in a lab or in the real world. If the mathematics make sense and the implementation is faithful to the model, the engine will work. Search engine effectiveness, on the other hand, cannot be tested within the formal model alone; for that one needs some experimental instrumentation and evaluation. We ask in this paper, how much and what kind of theoretical understanding the IR community has regarding IR effectiveness. The retrieval models do not cover these aspects – or at best, make strong implicit assumptions about it.

The goals of a research area may be classified as (a) theoretical understanding, (b) empirical description, prediction and explanation, and (c) technology development in the domain of interest. Much of research in IR is driven by a technological interest of developing tools for information access. However, technological interest becomes blind if not nurtured by the other goals. [6]

Reflecting this, the motivation for the present paper is that the ultimate goal of information retrieval is to support humans to better access information in order to better carry out their task. How well does IR effectiveness, measured at the output of search engines, reflect this? If IR effectiveness does not directly translate to better human information access, we risk turning means to ends with unfortunate consequences.

In the present paper, we make three points: (1) The Cranfield style of IR evaluation seems to lose power when one looks at human instead of system performance. (2)

Searchers using IR systems in real-life use rather short queries, which individually often have poor performance. However, when used in sessions, they may be surprisingly effective. The searcher's strategies have not been sufficiently described and cannot therefore be properly understood, supported nor evaluated. (3) Searchers in real-life seek to optimize the entire search process, not just result quality. Evaluation of output alone is insufficient to explain searcher behavior.

In Section 2, we review some past research on (non-formal) IR theory and introduce some concepts for discussing research approaches or paradigms. Section 3 discusses the limitations of the Cranfield approach in the light of recent empirical evidence. Section 4 takes a look at real-life IR based on sessions of short queries and argues that the Cranfield approach can be extended in this direction. Section 5 proposes a more holistic approach to IR evaluation based on searcher costs and efforts as well as output quality. Section 6 contains conclusions.

2 Past Analyses of IR Research

There are several introductions to approaches in IR research. For example, [6] reviewed three major approaches: systems-oriented IR, user-oriented IR and cognitive IR approaches. Järvelin [9] discussed the models and theories of systems-oriented and cognitive IR approaches. There is a dominant model for systems-oriented IR research, the Cranfield evaluation approach based on test collections. The other two major approaches do not have such dominant models.

Saracevic discussed critically evaluation in IR research and called for the integration of user-oriented and system-oriented IR research [13]. He criticized the sole use of relevance-based measures in evaluation and called for proper measures at the levels of users and uses, markets and products, and social impacts.

Ellis [3] questioned the applicability of the results of the Cranfield approach to operational systems due to validity problems in performance evaluation. He pointed out that the approach abstracts a mechanical component out of human interaction with texts at the cost of not being able to handle problems at the searcher level.

3 The Focus and Limits of Cranfield Style of IR Evaluation

3.1 The Cranfield Framework

Figure 1, center and left, represents the essence of Cranfield style of IR evaluation – the core components of experimental designs. In the unshaded center area there are documents/collections, requests/queries, and results; and core processes of representation and matching with some feedback. Their interaction is however laborious to study in real-life. We therefore want to move the components into a lab, and incorporate the shaded area, the necessary lab instrumentation. Using the instrumentation: standard collections, search requests, relevance assessments, and evaluation proce-

dures and metrics, enables us to effectively evaluate IR techniques and compare the results. Prototypically, the context and the user are excluded in experiments.

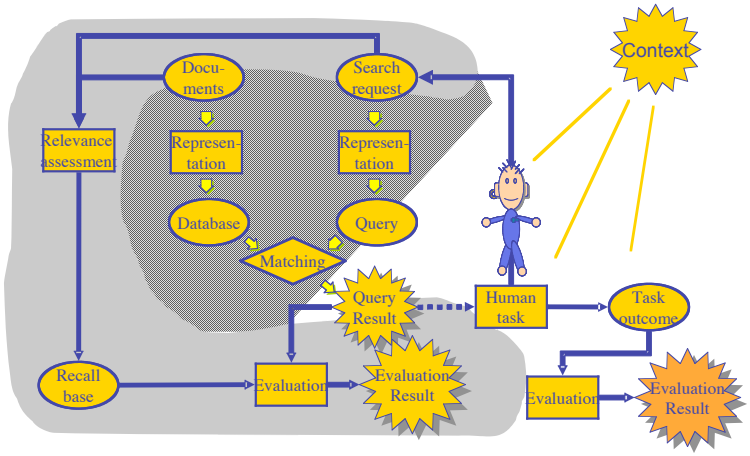

Fig. 1. Evaluation by human performance (extended from [9]).

3.2 Limitations of the Cranfield Approach: WYDSIWYDU

Cranfield style of IR research develops techniques for finding relevant documents. The quality of the techniques is usually measured through recall and precision of the output, or metrics based on these (e.g. MAP). The studies seek to explain the variation of output quality. The independent variables are the use or non-use of various IR techniques and the controlled ones the test collections, topics, assessments. [9]

The Cranfield framework has indispensable benefits that have lead to great progress. Standardization of experimental designs facilitates comparison of findings. One should not, however, be blinded by this success. A study designed within the Cranfield framework cannot claim anything about external variables: WYDSIWYDU – *what you don't see is what you don't understand*. These cover the tasks and searchers supported, relevance assessment, interface functionalities, and the actual search processes. There is mounting evidence that we may not be able to improve human performance by further improving traditional retrieval effectiveness.

3.3 Human Performance

Several recent studies have suggested that using a better search system may not always lead to improvements in task outcomes. Note here that we are stepping out of the lab, measuring something that essentially lies outside – the right side of Figure 1.

Allan and colleagues [1] studied searcher productivity in a passage-based question answering task. User performance improved significantly given a system performance improvement (from bpref of 50) whereafter system performance improvements did not yield a significant user performance change.

Turpin and Scholer [15] studied user performance on simple web search tasks, considering the time that a user takes to find a relevant document, and the number of relevant documents that a user can find within 5 minutes. This was studied across search systems operating at MAP in the range of 0.55 to 0.95. Results indicated that MAP level has no significant relationship with the time taken to find the first answer, while there was a weak relationship with the recall-oriented task.

Smith and Kantor [14] also explored the relation of system performance to search behavior. Their test subjects each completed several searches using either a standard system or a degraded system. Searchers using degraded systems were *as successful* as those using the standard one, regarding the quality of documents found and the time taken to achieve this. However, searchers using degraded systems *altered their behavior*, making significantly more queries and examining shorter lists.

Huuskonen and Vakkari [5] studied the connection of searching features to task outcome and found and found very few and vague connections between work task result quality and the system/searching variables.

These studies suggest that if one extends the Cranfield framework toward the human tasks, it loses strength. The main dependent variable, traditional IR effectiveness, is only weakly related to human task performance. Consequently, typical IR variables – IR techniques – do not explain the variation in the human task. Further, if the effect of the query result, measured through recall-precision metrics, is only weakly connected to human task performance, then:

- no experimentation with retrieval models will change the situation;
- no variation of evaluation metrics will change this if the metrics remain traditional;
- we need, in addition to result metrics, metrics for the process and the outcome.

4 Interaction in Sessions

Much IR research is based on batch mode experiments where a topic is automatically converted to a single multi-word query, which is then run against the database using some search engine. In real-life, searchers use very short queries but may try out multiple queries in a session. They also behave individually during search sessions. Their information needs may initially be muddled and change during the search process; they may learn as the session progresses, or switch focus. The initial query formulation may not be optimal and the searchers may need to try out different wordings. [10]

Real-life searchers often prefer short queries and avoid excessive browsing. Jansen and colleagues [8] analyzed transaction logs of thousands of queries posed to a Web search engine. The average query length was 2.21 keywords. Less than 4 % of the queries had more than 6 terms. They also observed that most users did not access results past the first page. Therefore real life sessions often consist of sequences of short queries. The data in Table 1 reflect these findings.

The data for Table 1 come from an empirical, interactive IR study [10]. Thirty domain experts each completed the same four realistic search tasks A – D simulating a need for specific information required to make a decision in a short time frame. Each task formed a session. The data show great variability between the tasks along various variables. On average, there were 2.5 queries per session and 2.4 unique keys per

session, and each query had two keys and 0.9 filters (a geographic, document type or other condition). Only 10 among the 60 sessions employed four or more unique search keys. These searchers were precision-oriented, i.e., they quit searching soon after finding one or a few relevant documents.

Table 1. Real-life session statistics based on 15 sessions for Tasks A-D (N=60) sessions.

Variable	A	B	C	D	Tot
Tot # queries per task	25	59	28	40	152
Avg queries in session	1.7	3.9	1.9	2.7	2.5
Avg # keys per session	1.5	3.9	1.9	2.2	2.4
Avg # keys per query	1.4	2.4	1.8	2.0	2.0
Avg # filters per query	1.2	1.1	0.8	0.7	0.9
S1 frequency	11	3	4	3	21
S2 frequency	2	4	3	4	13
S3 frequency	4	13	11	10	38
S1-S3 frequency sum	17	20	18	17	72

The four bottoms lines report the *frequency of the query strategies* discussed below. The strategies S1, S2, and S3 were identified in Table 1 session data through a secondary analysis [11]. Strategy S1 consisted of individual words used alone as queries. If the first word was unsuccessful, another was tried instead. S1 was employed 21 times in the 60 sessions of Table 1. Strategy S2 is based on incremental query extension: a searcher starts with a one word query. If it is not successful, (s)he extends the query by another word, by a third word, etc. S2 was employed in 13 times of the 60 sessions of Table 1. Strategy S3 is based on three word queries where the first two are fixed and the third one varied. S3 was employed in 38 of the 60 sessions. The total number of identified strategies (72) exceeds the number of sessions (60) because more than one strategy was employed in some sessions. For completeness, strategy S4 is defined as a full multiword query based on a test topic – no-one used it in the empirical data.

Keskustalo and colleagues [11] used the strategies S1-S4 in a simulation experiment based on TREC 7 and 8 test collections (528155 documents) and 41 topics for which graded relevance assessments were available. The retrieval system was *Lemur*. The authors used real test persons to suggest keywords for queries of various lengths. These were then used to construct simulated sessions following the Strategies S1-S4 with an interest in finding whether sessions based on simple queries and Strategies S1-S3 are competitive with verbose individual queries using Strategy S4. Sessions of five queries were used for strategies S1-S3 and the search task was to find one highly relevant document, which is a frequently used task in interactive IR experiments. For each of 5 queries in S1-S3, only the first result page was examined. For S4, the top-50 was examined in lots of ten results for compatibility.

Taken individually, the queries in sessions of Strategies S1-S3 often had poor effectiveness (e.g. measured by MAP). However, session effectiveness of S1-S3 was considerably higher (Table 2). The average page of success tells which page, on the average, contained the first highly relevant document. Based on stringent relevance criteria, S1 is 20-34 percent units weaker than strategies S2-S4 in its success rate. Strategies S2-S3 are only 8-13 % units below S4. In all cases, the first highly relevant document is found, on average, by the second attempt (second page for S4). Accord-

ing to Friedman's test the differences between the strategies are highly significant ($p < 0.001$). In pairwise tests, S3 is not significantly different from S4 while S2-S4 are all significantly better than S1 ($p<0.01$).

Table 2. Effectiveness of session strategies S1-S4 for 41 topics as average page of success.

Variable	S1	S2	S3	S4
Avg successful page	1,82	1,73	1,52	1,42
Success rate	23	33	31	36
Success %	60,5	86,8	81,6	94,7

This study shows that sessions based on individually ineffective queries may be surprisingly effective. The findings motivate the observed real-life user behavior, which real users must have learned through experience. As few very simple attempts often lead to good enough results, there is no incentive to pay more effort.

5 Toward a Holistic View on IR

We believe that research on IR interaction is currently too exclusively focused on the quality of retrieval results. Early papers on IR evaluation had a comprehensive approach: Cleverdon and colleagues [2] and Salton [12] identified, among others, presentation issues and intellectual and physical user effort as important factors in IR evaluation, along with recall and precision as performance measures. Usability studies also have a comprehensive approach to costs and benefits of systems assessed [7].

Hersh [4] pointed out that the potential impact of an interactive IR system is determined in part by situational relevance, which is affected, among others, by the user's time pressure. Therefore only documents retrieved in the top ranks of results may be of interest. Järvelin and colleagues [10] extended the Discounted Cumulated Gain metric into a session-based evaluation metric (sDCG), which handles multiple query sessions and takes the searcher's effort (both scanning and query modification costs) indirectly into account through discounting factors.

While costs and benefits of interactive IR systems have been discussed in the literature, the same does not hold for current IR evaluation, which seems to focus on retrieval result quality and neglect searcher efforts. In interactive settings both (expected) costs and benefits affect searcher behavior and evaluation becomes biased if only result quality is considered. To avoid this problem, a cost/benefit model for interactive IR sessions is needed. It should incorporate at least the following cost/benefit factors in a typical search interface:

- Search key generation cost (K): the cost of producing each search key.
- Query execution cost (Q): the cost of giving a search and waiting for the result.
- Result scan cost (S): The cost of scanning each item in the result.
- Next page access cost (N): the cost of accessing the next results page for scanning.
- Relevant document gain (G): the benefit of identifying a relevant document.

A rough cost/benefit model assumes all the above costs linear per respective numbers of units, in the same value range (e.g. seconds), and additive. When this is made com-

mensurate with the relevant document gain, e.g. by a conversion factor between costs and gains, one may use the following function *SessionCBA* for evaluation:

$$SessionCBA(K,Q,S,N,G) = \alpha K + \beta Q + \delta S + \gamma N + \theta G \quad (1)$$

where α, β, δ, γ, and θ are constant unit costs/benefits of the above variables K, Q, S, N, and G. Note that traditional IR evaluation assumes $\alpha = \beta = \delta = \gamma = 0$ and $\theta > 0$, thus making $SessionCBA(K,Q,S,N,G) = \theta G$ and focusing on benefits at any cost. This can hardly be used to explain searcher behavior.

As an example, consider the case in Section 4 by [11]. Table 3 gives the expected number of search keys K, queries Q, scanned documents S, fetched next pages N, and found relevant documents G (one in each case) for each strategy.

Table 3. Cost-benefit features of Strategies S1-S4.

Strategy	K	Q	S	N	G
S1	8.6	3.5	35	0.0	1
S2	4.3	2.3	23	0.0	1
S3	7.3	2.4	24	0.0	1
S4	16.9	1.0	5	0.0	1

The cost-benefit features of Strategies S1-S3 are calculated based on the success statistics of 1^{st} – 5^{th} queries and on the assumption that, if none of them is successful (see Table 2), that the searcher would launch one more query represented by S4 containing 16.9 search keys. If the action would be just giving up without an answer after five unsuccessful attempts, the K column would have values 3.9, 2.6, 5.1, and 16.9 keys. This may happen, if target information is not very valuable.

Because we do not know the unit costs of K, Q, S, and N, we cannot directly identify an optimal strategy. One may still observe that if entering query words is costly and scanning the result cheap, S1-S3 are competitive, whereas in the opposite case S4 wins.

6 Discussion and Conclusion

Theoretical growth in a research area may incur from theory expansion (e.g., through new concepts), greater analytical power (through model building), improved empirical support, and proliferation of new hypotheses within the theory [16].

Section 3 discussed the Cranfield IR evaluation framework and its limitations in the light human task performance. Based on several critical studies we found evidence suggesting that the Cranfield style IR evaluation framework is weakly connected to human task performance. In the effort of making experiments controllable, the Cranfield approach may have crystallized a study design that weakly relates to the activity supported. No experimentation with IR techniques or traditional evaluation metrics will change the situation. To explain user performance theory expansion is necessary.

We then discussed interaction in real-life IR sessions and discussed three idealized real-life, session-based, retrieval strategies S1-S3 as alternatives to a long test query S4. A simulated interactive retrieval experiment showed that sessions using individually ineffective queries may be surprisingly effective. The findings motivate the ob-

served real-life user behavior, which real users must have learned through experience with IR systems. This suggests that greater analytical power is needed for understanding user behavior. Section 5 exemplifies that this can be achieved by more holistic modeling of both session costs and benefits for better empirical support.

In conclusion, there are risks in focusing wholly on IR tools without analyzing their real use contexts. One cannot understand their use, nor design them properly, without understanding at least minimally the information environment of their users.

Acknowledgement

This research was funded by the Academy of Finland grants #120996 and #124131.

References

1. Allan, J., Carterette, B., Lewis, J.: When will information retrieval be "good enough"? In: Proc. ACM SIGIR'05, pp. 433--440. ACM Press, New York (2005)
2. Cleverdon, C. Mills, L., Keen M.: Factors determining the performance of indexing systems, vol. 1 - design. Aslib Cranfield Research Project, Cranfield (1966)
3. Ellis, D.: Progress and problems in information retrieval. Library Assoc., London (1996)
4. Hersh, W.: Relevance and Retrieval Evaluation: Perspectives from Medicine. J. Amer. Soc. Inform. Sci. 45, 201--206 (1994)
5. Huuskonen, S., Vakkari, P.: Students' search process and outcome in Medline in writing an essay for a class on evidence based medicine. J. Documentat. 64, 287--303 (2008)
6. Ingwersen, P., Järvelin K.: The turn: Integration of information seeking and retrieval in context. Springer, Heidelberg (2005)
7. ISO Ergonomic requirements for office work with visual display terminals (VDTs), Part 11: Guidance on usability. ISO 9241-11:1998(E), (1998)
8. Jansen, B.J., Spink, A., Saracevic, T.: Real Life, Real Users, and Real Needs: A Study and Analysis of User Queries on the Web. Inform. Proc. Manag. 36, 207-227 (2000)
9. Järvelin, K.: An Analysis of Two Approaches in Information Retrieval: From Frameworks to Study Designs. J. Amer. Soc. Inform. Sci. 58, 971--986 (2007)
10. Järvelin, K., Price, S.L., Delcambre, L.M.L., Nielsen, M.L.: Discounted Cumulated Gain Based Evaluation of Multiple-Query IR Sessions. In: Proc. ECIR'08. LNCS, vol. 4956, pp. 4--15. Springer, Heidelberg (2008)
11. Keskustalo, H. & Järvelin, K. & Pirkola, A. & Sharma, T. & Lykke Nielsen, M.: Test Collection-Based IR Evaluation Needs Extension Toward Sessions - A Case Study of Extremely Short Queries. Proc. AIRS 2009. LNCS, to appear. Springer, Heidelberg (2009)
12. Salton, G.: Evaluation Problems in Interactive Information Retrieval. Inform. Stor. Retr. 6, 29--44 (1970)
13. Saracevic, T.: User lost: reflections on the past, future, and limits of information science. ACM SIGIR Forum 31(2), 16--27 (1997)
14. Smith, C.L., Kantor, P.B.: User Adaptation: Good Results from Poor Systems. In: Proc. ACM SIGIR'08, pp. 147--154. ACM Press, New York (2008)
15. Turpin, A., Scholer, F.: User performance versus precision measures for simple search tasks. In: Proc. ACM SIGIR'06, pp.11--18. ACM Press, New York (2006)
16. Wagner, D., Berger, J., Zeldith, M.: A working strategy for constructing theories. In: G. Ritzer (Ed.), Metatheorizing, pp. 107-123. Sage (1992)