A Plan for Making Information Retrieval Evaluation Synonymous with Human Performance Prediction

Mark D. Smucker
Department of Management Sciences
University of Waterloo
msmucker@uwaterloo.ca

ABSTRACT

Today human performance on search tasks and information retrieval evaluation metrics are loosely coupled. Instead, information retrieval evaluation should be a direct prediction of human performance rather than a related measurement of ranked list quality. We propose a TREC track or other group effort that will collect a large amount of human usage data on search tasks and then measure participating sites' ability to develop models that predict human performance given the usage data. With models capable of accurate human performance prediction, automated information retrieval evaluation should become an even better tool for driving the future of information retrieval research.

1. INTRODUCTION

In many respects, we believe that the future of information retrieval (IR) evaluation has already been written. In 1973, Cooper [8, 9] wrote a two-part paper outlining what he believed the evaluation of IR should be. In part 1, Cooper presented his "naive evaluation methodology" that held that IR effectiveness should be based on the users' personal utility gained from using an IR system. In part 2, Cooper put forth a possible plan of research that would establish ways to approximate this utility and in particular proposed *validation experiments* to measure the ability of an evaluation method to predict utility. With the rapid changes in computing and the fields of IR and human computer interaction (HCI) it is not too surprising that Cooper's vision was not quickly realized.

In 2009, we see a building momentum for adoption of these ideas but the majority of IR evaluations still focus only on measuring ranking quality with variants of precision and recall that are only loosely predictive of utility [2, 3, 4, 13, 26, 27]. In other words, today's IR researchers tend to evaluate IR systems much as was done prior to Cooper's proposal. In this paper, we renew Cooper's call for the future of IR evaluation and outline a plan to help the IR community move toward evaluation focused on *human performance prediction*.

2. REALIZING COOPER'S VISION

The Cranfield or "batch mode" style of evaluation has been a corner stone of IR progress for over 40 years and serves a complementary role to manual user studies. Cranfield style evaluation takes a ranked list of documents produced by a retrieval system in response to a query and evaluates the list by using a pre-existing set of relevance judgments.

A consistent criticism of the Cranfield style of evaluation is that it does not reflect the wide range of user behavior observed with interactive IR systems.

An important step towards realizing Cooper's vision was taken by Dunlop [11], who in 1997 made a case for the following ideas:

- Evaluation should be *predictive* of user performance.
- Evaluation should concern itself with both the user interface and the underlying retrieval engine.
- Evaluation should measure the time required for users to satisfy their information needs.

Whereas Cooper proposed to measure users' subjective utility, Dunlop examined performance with plots of time vs. number of relevant documents found — a measure inspired by Cooper's expected search length [7]. To make predictions of user performance, Dunlop built user models utilizing HCI methods developed in the decades following Cooper's proposal. Dunlop left as future work the validation of his predictions, i.e. a Cooper validation experiment.

While human performance is not always the same as users' subjective utility, we see Dunlop's ideas in combination with Cooper's validation experiments as the next step towards realizing Cooper's vision and the future of IR evaluation.

3. A BUILDING MOMENTUM

Dunlop's evaluation methodology is still a batch-mode evaluation that relies on a Cranfield style test collection. As Lin and Smucker [20] explain, the Cranfield style of evaluation can be seen as a form of automated usability [14] where the evaluation consists of some hypothetical user interface and a model of user behavior over that interface.

In the case of a Cranfield style evaluation, the hypothetical user interface allows for a query and display of a ranked list of results. The Cranfield style user model assumes the user will examine the results in rank order at a uniform rate and then stop at the end of the ranked list.

Dunlop extended the Cranfield style of evaluation to allow for different user interfaces and to utilize user models that predicted the time to examine the displayed ranked lists.

While not making time-based predictions and utilizing simple user models, several other researchers have also aimed to simulate the use of interactive IR systems [1, 19, 20, 25,

SIGIR Workshop on the Future of IR Evaluation, July 23, 2009, Boston.
.

28]. Azzopardi [5] provides a useful discussion on the use of examined document sequences for evaluation of interactive IR as utilized by Aalbersberg [1] and others.

At the same time, work has been occurring that has in effect kept the hypothetical user interface fixed to the simple single query, single results paradigm but has aimed to incorporate different user models. Some of this work incorporates a user model into the retrieval metric with the focus on modeling when the user stops examining documents in the ranked list [7, 10, 15, 22].

Another body of work has utilized HCI user modeling techniques (c.f. Dunlop) to IR and IR-related tasks [6, 12, 17, 21, 23, 24]. In many of these cases, the simulations are compared to actual human studies to determine if the user model accurately reflects human performance.

Recently, Keskustalo et al. [18] have taken a significant step forward in simulating human search behavior with an evaluation methodology that examines and simulates query reformulation.

4. OUTLINE OF PLAN

We propose a TREC track or other group effort that defines a canonical search user interface (UI) and collects a large amount of user behavior on TREC-styled ad-hoc search topics. The aim of this effort is to evaluate different evaluation methods on their ability to predict actual human search behavior and performance.

We are only proposing to move IR evaluation in one direction: better prediction of human performance. There are many dimensions to IR evaluation and we do not aim to change the current accepted practices in these other dimensions. For example, we think the task should largely remain a searching of newswire documents, the saving of relevant documents, and the using of an interface that consists of a search box and 10 top ranked results with query-biased snippets.

This effort would in effect create an interaction pool [16] with possibly many participants plugging different retrieval engines into the canonical UI. An attempt would need to be made to collect as much relevant interaction data as possible (queries, clicks, keystrokes, mouse movement, eye tracking, server response times, time documents are saved, etc.).

In summary, we would collect real user data telling us when relevant information is discovered. This data will give us the means to train and test models of human performance prediction — a possible TREC track evaluation of evaluation methods.

5. CONCLUSION

Will it be easy to collect enough user interaction data to make it possible for new evaluation techniques to be created and tested on their ability to predict human search performance? No, but we believe it is preferable to directly predict human performance rather than continue in the current cycle of creating retrieval metrics and then post-hoc testing their predictive ability with expensive user studies.

6. REFERENCES

[1] I. J. Aalbersberg. Incremental relevance feedback. In *SIGIR'92*, pp. 11–22.

[2] A. Al-Maskari, M. Sanderson, and P. Clough. The relationship between IR effectiveness measures and user satisfaction. In *SIGIR'07*, pp. 773–774.

[3] A. Al-Maskari, M. Sanderson, P. Clough, and E. Airio. The good and the bad system: does the test collection predict users' effectiveness? In *SIGIR'08*, pp. 59–66.

[4] J. Allan, B. Carterette, and J. Lewis. When will information retrieval be "good enough"? In *SIGIR'05*, pp. 433–440.

[5] L. Azzopardi. Towards evaluating the user experience of interactive information access systems. In *SIGIR'07 Web Information-Seeking and Interaction Workshop.*

[6] E. H. Chi, P. Pirolli, K. Chen, and J. Pitkow. Using information scent to model user information needs and actions and the web. In *CHI'01*, pp. 490–497.

[7] W. S. Cooper. Expected search length: A single measure of retrieval effectiveness based on the weak ordering action of retrieval systems. *Am. Doc.*, 19(1):30–41, Jan 1968.

[8] W. S. Cooper. On selecting a measure of retrieval effectiveness. *JASIS*, 24(2):87–100, Mar/Apr 1973.

[9] W. S. Cooper. On selecting a measure of retrieval effectiveness: Part ii. implementation of the philosophy. *JASIS*, 24(6):413–424, Nov/Dec 1973.

[10] A. P. de Vries, G. Kazai, and M. Lalmas. Tolerance to irrelevance: A user-effort oriented evaluation of retrieval systems without predefined retrieval unit. In *RIAO'04*, pp. 463–473.

[11] M. D. Dunlop. Time, relevance and interaction modelling for information retrieval. In *SIGIR'97*, pp. 206–213.

[12] W.-T. Fu and P. Pirolli. Snif-act: a cognitive model of user navigation on the world wide web. *Hum.-Comput. Interact.*, 22(4):355–412, 2007.

[13] W. Hersh, A. Turpin, S. Price, B. Chan, D. Kramer, L. Sacherek, and D. Olson. Do batch and user evaluations give the same results? In *SIGIR'00*, pp. 17–24.

[14] M. Y. Ivory and M. A. Hearst. The state of the art in automating usability evaluation of user interfaces. *ACM Computing Surveys*, 33(4):470–516, 2001.

[15] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of ir techniques. *TOIS*, 20(4):422–446, 2002.

[16] H. Joho, R. Villa, and J. Jose. Interaction pool: Towards a user-centered test collection. In *SIGIR'07 Web Information-Seeking and Interaction Workshop.*

[17] M. T. Keane, M. O'Brien, and B. Smyth. Are people biased in their use of search engines? *Commun. ACM*, 51(2):49–52, 2008.

[18] H. Keskustalo, K. Järvelin, T. Sharma, and M. L. Nielsen. Test collection-based IR evaluation needs extension toward sessions: A case of extremely short queries. In *AIRS'09*.

[19] J. Lin. User simulations for evaluating answers to question series. *IPM*, 43(3):717–729, 2007.

[20] J. Lin and M. D. Smucker. How do users find things with PubMed? Towards automatic utility evaluation with user simulations. In *SIGIR'08*, pp. 19–26.

[21] C. S. Miller and R. W. Remington. Modeling information navigation: implications for information architecture. *Hum.-Comput. Interact.*, 19(3):225–271, 2004.

[22] A. Moffat and J. Zobel. Rank-biased precision for measurement of retrieval effectiveness. *TOIS*, 27(1):1–27, 2008.

[23] V. A. Peck and B. E. John. Browser-soar: a computational model of a highly interactive task. In *CHI'92*, pp. 165–172.

[24] P. Pirolli. *Information Foraging Theory.* 2007.

[25] M. D. Smucker and J. Allan. Find-similar: Similarity browsing as a search tool. In *SIGIR'06*, pp. 461–468.

[26] A. Turpin and F. Scholer. User performance versus precision measures for simple search tasks. In *SIGIR'06*, pp. 11–18.

[27] A. H. Turpin and W. Hersh. Why batch and user evaluations do not give the same results. In *SIGIR'01*, pp. 225–231.

[28] R. W. White, J. M. Jose, C. J. van Rijsbergen, and I. Ruthven. A simulated study of implicit feedback models. In *ECIR'04*, pp. 311–326.