The fault, dear researchers, is not in Cranfield, But in our metrics, that they are unrealistic.

Mark D. Smucker
Department of Management Sciences
University of Waterloo, Canada
mark.smucker@uwaterloo.ca

Charles L. A. Clarke
School of Computer Science
University of Waterloo, Canada
claclark@plg.uwaterloo.ca

1. INTRODUCTION

As designers of information retrieval (IR) systems, we need some way to measure the performance of our systems. An excellent approach to take is to directly measure actual user performance either in situ or in the laboratory [12]. The downside of live user involvement is the prohibitive cost if many evaluations are required. For example, it is common practice to sweep parameter settings for ranking algorithms in order to optimize retrieval metrics on a test collection. The Cranfield approach to IR evaluation provides low-cost, reusable measures of system performance.

Cranfield-style evaluation frequently has been criticized as being too divorced from the reality of how users search, but there really is nothing wrong with the approach [18]. The Cranfield approach effectively is a simulation of IR system usage that attempts to make a prediction about the performance of one system vs. another [15].

As such, we should really be thinking of the Cranfield approach as the application of models to make predictions, which is common practice in science and engineering. For example, physics has equations of motion. Civil engineering has models of concrete strength. Epidemiology has models of disease spread. Etc. In all of these fields, it is well understood that the models are simplifications of reality, but that the models provide the ability to make useful predictions.

Information retrieval's predictive models are our evaluation metrics.

The criticism of system-oriented IR evaluation should be redirected. The problem is not with Cranfield — which is just another name for making predictions given a model — the problem is with the metrics.

We believe that rather than criticizing Cranfield, the correct response is to develop better metrics. We should make metrics that are more predictive of human performance. We should make metrics that incorporate the user interface and realistically represent the variation in user behavior. We should make metrics that encapsulate our best understanding of search behavior.

In popular parlance, we should bring solutions, not problems, to the system-oriented IR researcher. To this end, we have developed a new evaluation metric, time-biased gain (TBG), that predicts IR system performance in human terms of the expected number of relevant documents to be found by a user [16].

2. TIME-BIASED GAIN

HCI has a long history of automated usability evaluation [10], and indeed, so does IR. Cleverdon designed the Cranfield 2 study carefully in terms of a specific type of user and how this type of user would define relevance [8, p. 9]. Taken together, a test collection (documents, topics, relevance judgments) and an evaluation metric allow for the simulation of a user with different IR systems.

Järvelin and Kekäläinen produced a significant shift in evaluation metrics with their introduction of cumulated gain-based measures [11]. The cumulated gain measures are explicitly focused on a model of a user using an IR system. As long as the user continues to search, the user can continue to increase their gain. The common notion of gain in IR evaluation is the relevant document, but gain can be anything we would like to define it to be.

Cumulated gain can be plotted vs. time to produce a gain curve and compare systems. The curve that rises higher and faster than another curve is the preferred curve. While we can plot gain curves of one system vs. another, it is well-known that users do not endlessly search; different users stop their searches at different points in time for a host of reasons. Given a probability density function $f(t)$ that gives the distribution of time spent searching, we can compute the expected gain as follows:

$$E[G(t)] = \int_0^\infty G(t) f(t) dt, \quad (1)$$

where $G(t)$ is the cumulated gain at time t. Equation 1 represents *time-biased gain* in its general form, i.e. time-biased gain is the expected gain for a population of users.

While it is natural for us to talk about cumulated gain over time, the traditional cumulated gain measures have substituted document rank for time and implicitly model a user that takes the same amount of time to evaluate each and every document. By making time a central part of our metric, we gain the ability to more accurately model behavior. For example, in a document retrieval system, longer documents will in general take users longer to evaluate, and if the retrieval system presents results with document summaries (snippets), we know that users can use summaries to speed the rate at which they find relevant information [14].

Another significant advantage of using time directly in our retrieval metric is that we now make testable predictions of human performance. Our predictions are in the same units as would be obtained as part of a user study. To our knowledge, this alignment between the units of Cranfield-style metrics and user study metrics has not previously existed.

Time-biased gain in the form of Equation 1 makes no mention of ranked lists of documents, for it is a general purpose description of users using an IR system over time. To produce a metric suitable for use in evaluating ranked lists, we followed a process common to development of new simulations [3]:

1. Creation of model.
2. Calibration of model.
3. Validation of model.

Our first step in model creation was to adopt the standard model of a user that works down a result list and move Equation 1 to a form common to cumulated gain measures:

$$\sum_{k=1}^{\infty} g_k D(T(k)), \tag{2}$$

where g_k is the gain at rank k, $T(k)$ is the expected time it takes a user to reach rank k, and $D(t)$ is the fraction of the population that survives to time t and is called the decay function.

Our model for the time it takes a user to reach rank k, $T(k)$, takes into consideration a hypothetical user interface that presents results to the user in the form of document summaries. A click on a document summary takes the user to the full document. We model both the probabilities of clicking on summaries given their NIST relevance and the probability of then judging a viewed full document as relevant. We separately model the time to view summaries and full documents. For the time spent on a full document, we modeled longer documents taking longer with an additional constant amount of spent. We treated duplicate documents as zero length documents. We then calibrated $T(k)$ using data from a user study, and finally we validated that our $T(k)$ provided a reasonable fit to the user study data. Likewise, we modeled $D(t)$ as exponential decay fit to a search engine's log data.

In contrast, older evaluation metrics such as mean average precision [19, p. 59] cannot be calibrated and have only been validated after their creation. For example, the work of Hersh and Turpin [9] is likely the first attempt to validate a metric (average precision). Many recent metrics can be calibrated to actual user behavior [4, 5, 7, 17, 20, 21], but their calibration and validation often come after their release and adoption.

3. CONCLUSION

The Cranfield approach to IR evaluation is merely another name for the development and use of predictive models, which is a fundamental part all science and engineering fields. In particular, IR evaluation fits nicely into the framework of simulation where models are created, calibrated, and validated before being used to make predictions. We have presented time-biased gain as an example of what we believe the correct direction is for IR system evaluation. We are not the only ones to be working on better metrics or taking a simulation based approach [2, 13], and others also consider time an important part of evaluation [1, 6].

Our position is that system-oriented IR research is user-oriented IR research given its use of evaluation metrics that model users. If HCIR researchers can produce better models than exist today — by better, we mean more predictive of human performance — then we can help system development to focus on changes that help users better search.

4. ACKNOWLEDGMENTS

This work was supported in part by the NSERC, in part by GRAND NCE, in part by Google, in part by Amazon, in part by the facilities of SHARCNET, and in part by the University of Waterloo. Any opinions, findings and conclusions or recommendations expressed in this material are the authors' and do not necessarily reflect those of the sponsors.

5. REFERENCES

[1] L. Azzopardi. Usage based effectiveness measures: monitoring application performance in information retrieval. *CIKM*, pages 631–640, 2009.

[2] L. Azzopardi, K. Järvelin, J. Kamps, and M. D. Smucker. Report on the SIGIR 2010 workshop on the simulation of interaction. *SIGIR Forum*, 44:35–47, January 2011.

[3] J. Banks, J. S. Carson II, B. L. Nelson, and D. M. Nicol. *Discrete-Event System Simulation*. Prentice Hall, 5th edition, 2010.

[4] B. Carterette, E. Kanoulas, and E. Yilmaz. Simulating simple user behavior for system effectiveness evaluation. In *CIKM*, pages 611–620, 2011.

[5] O. Chapelle, D. Metlzer, Y. Zhang, and P. Grinspan. Expected reciprocal rank for graded relevance. In *CIKM*, pages 621–630, Hong Kong, 2009.

[6] M. D. Dunlop. Time, relevance and interaction modelling for information retrieval. In *SIGIR*, pp. 206–213. 1997.

[7] G. Dupret. Discounted cumulative gain and user decision models. In *Proceedings of the 18th international conference on String processing and information retrieval*, SPIRE'11, pages 2–13, Berlin, Heidelberg, 2011. Springer-Verlag.

[8] D. Harman. *Information Retrieval Evaluation*. Morgan & Claypool, 2011.

[9] 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*, pages 17–24. ACM, 2000.

[10] 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.

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

[12] D. Kelly. *Methods for Evaluating Interactive Information Retrieval Systems with Users*, volume 3. Foundations and Trends in Information Retrieval, 2009.

[13] 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*, pp. 63–74, 2009.

[14] R. Khan, D. Mease, and R. Patel. The impact of result abstracts on task completion time. In *Workshop on Web Search Result Summarization and Presentation, WWW'09*, 2009.

[15] J. Lin and M. D. Smucker. How do users find things with PubMed? Towards automatic utility evaluation with user simulations. In *SIGIR'08*, pages 19–26. ACM, 2008.

[16] M. D. Smucker and C. L. A. Clarke. Time-based calibration of effectiveness measures. In *SIGIR*, 10 pages, 2012.

[17] A. Turpin, F. Scholer, K. Järvelin, M. Wu, and J. S. Culpepper. Including summaries in system evaluation. In *SIGIR'09*, pages 508–515. ACM, 2009.

[18] E. M. Voorhees. I come not to bury Cranfield, but to praise it. In *HCIR'09*, pages 13–16, 2009.

[19] E. M. Voorhees and D. K. Harman, editors. *TREC*. MIT Press, 2005.

[20] E. Yilmaz, M. Shokouhi, N. Craswell, and S. Robertson. Expected browsing utility for web search evaluation. In *CIKM*, pages 1561–1564, Toronto, 2010.

[21] Y. Zhang, L. A. Park, and A. Moffat. Click-based evidence for decaying weight distributions in search effectiveness metrics. *Information Retrieval*, 13:46–69, February 2010.