Resolving the Battle Royale between Information Retrieval and Information Science

Daniel Tunkelang
Endeca
dt@endeca.com

ABSTRACT

We propose an approach to help resolve the "battle royale" between the information retrieval and information science communities. The information retrieval side favors the Cranfield paradigm of batch evaluation, criticized by the information science side for its neglect of the user. The information science side favors user studies, criticized by the information retrieval side for their scale and repeatability challenges. Our approach aims to satisfy the primary concerns of both sides.

Categories and Subject Descriptors

H.1.2 [Human Factors]: Human information processing.

H.3.3 [Information Systems]: Information Search and Retrieval - Information Filtering, Retrieval Models

H.5.2 [Information Systems]: Information Interfaces and Presentation - User Interfaces

General Terms

Design, Experimentation, Human Factors

Keywords

Information science, information retrieval, information seeking, evaluation, user studies

1. INTRODUCTION

Over the past few decades, a growing community of researchers has called for the information retrieval community to think outside the Cranfield box. Perhaps the most vocal advocate is Nick Belkin, whose "grand challenges" in his keynote at the 2008 European Conference on Information Retrieval [1] all pertained to the interactive nature of information seeking he claims the Cranfield approach neglects. Belkin cited similar calls to action going back as far as Karen Spärck Jones, in her 1988 acceptance speech for the Gerald Salton award [2], and again from Tefko Saracevic, when he received the same award in 1997 [3]. More recently, we have the Information Seeking and Retrieval research program proposed by Peter Ingwersen and Kalervo Järvelin in *The Turn*, published in 2005 [4].

Information Seeking Support Systems Workshop, June 26–27, 2008, Chapel Hill, North Carolina, USA.

2. IMPASSE BETWEEN IR AND IS

Given the advocacy of Belkin and others, why hasn't there been more progress? As Ellen Voorhees noted in defense of Cranfield at the 2006 Workshop on Adaptive Information Retrieval, "changing the abstraction slightly to include just a bit more characterization of the user will result in a dramatic loss of power or increase in cost of retrieval experiments" [5]. Despite user studies that have sought to challenge the Cranfield emphasis on batch information retrieval measures like mean average precision—such as those of Andrew Turpin and Bill Hersh [6]—the information retrieval community, on the whole, remains unconvinced by these experiments because they are smaller in scale and less repeatable than the TREC evaluations.

As Tefko Saracevic has said, there is a "battle royale" between the information retrieval community, which favors the Cranfield paradigm of batch evaluation despite its neglect of the user, and the information science community, which favors user studies despite their scale and repeatability challenges [7]. How do we move forward?

3. PRIMARY CONCERNS OF IR AND IS

Both sides have compelling arguments. If an evaluation procedure is not repeatable and cost-effective, it has little practical value. Nonetheless, it is essential that an evaluation procedure measure the interactive nature of information seeking.

If we are to find common ground to resolve this dispute, we need to satisfy the primary concerns of both sides:

- Real information seeking tasks are interstice, so the results of the evaluation procedure must be meaningful in an interactive context.
- The evaluation procedure must be repeatable and cost-effective.

In order to move beyond the battle royale and resolve the impasse between the IR and IS communities, we need to address both of these concerns.

4. PROPOSED APPROACH

A key point of contention in the battle royale is whether we should evaluate systems by studying individual users or measuring system performance against test collections.

The short answer is that we need to do both. In order to ground the results of evaluation in realistic contexts, we need to conduct user studies that relate proposed measures to success in interactive information seeking tasks. Otherwise, we optimize under the artificial constraint that a task involves only a single user query.

Such an approach presumes that we have a characterization of information seeking tasks. This characterization is an open problem that is beyond the scope of this position paper but has been addressed by other information seeking researchers, including Ingwersen and Järvelin [4]. We presume access to a set of tasks that, if not exhaustive, at least applies to a valuable subset of real information seeking problems.

Consider, as a concrete example, the task of a researcher who, given a comprehensive digital library of technical publications, wants to determine with confidence whether his or her idea is novel. In other words, the researcher want to either discover prior art that anticipates the idea, or to state with confidence that there is no such art. Patent inventors and lawyers performing e-discovery perform analogous tasks. We can measure task performance objectively as a combination of accuracy and efficiency, and we can also consider subject measures like user confidence and satisfaction. Let us assume that we are able to quantify a task success measure that incorporates these factors.

Given this task and success measure, we would like to know how well an information retrieval system supports the user performing it. As the information scientists correctly argue, user studies are indispensable. But, as we employ user studies to determine which systems are most helpful to users, we need to go a step further and correlate user success to one or more system measures. We can then evaluate these system measures in a repeatable, cost-effective process that does not require user involvement.

For example, let us hypothesize that mean average precision (MAP) on a given TREC collection is such a measure. We hypothesize that users pursuing the prior art search task are more successful using a system with higher MAP than those using a system with lower MAP. In order to test this hypothesis, we can present users with a family of systems that, insofar as possible, vary only in MAP, and see how well user success correlates to the system's MAP. If the correlation is strong, then we validate the utility of MAP as a system measure and invest in evaluating systems using MAP against the specified collection in order to predict their utility for the prior art task.

The principle here is a general one, and can even be used not only to compare different algorithms, but also to evaluate more sophisticated interfaces, such as document clustering [8] or faceted search [9]. The only requirement is that we hypothesize and validate system measures that correlate to user success.

5. WEAKNESSES OF APPROACH

Our proposed approach has two major weaknesses.

The first weakness is that, in a realistic interactive information retrieval context, distinct queries are not independent. Rather, a typical user executes a sequence of queries in pursuit of an information need, each query informed by the results of the previous ones.

In a batch test, we must decide the query sequence in advance, and cannot model how the user's queries depend on system response. Hence, we are limited to computing measures that can be evaluated for each query independently. Nonetheless, we can choose measures which correlate to effectiveness in realistic settings. Hopefully these measures are still meaningful, even when we remove the test queries from their realistic context.

The second challenge is that we do not envision a way to compare different interfaces in a batch setting. It seems that testing the relative merits of different interfaces requires real—or at least simulated—users.

If, however, we hold the interface constant, then we can define performance measures that apply to those interfaces. For example, we can develop standardized versions of well-studied interfaces, such as faceted search and clustering. We can then compare the performance of different systems that use these interfaces, e.g., different clustering algorithms.

6. AN ALTERNATIVE APPROACH

An alternative way to tackle the evaluation problem leverages the "human computation" approach championed by Luis Von Ahn [10]. This approach uses "games with a purpose" to motivate people to perform information-related tasks, such as image tagging and optical character recognition (OCR).

A particularly interesting "game" in our present context is Phetch, in which in which one or more "Seekers" compete to find an image based on a text description provided by a "Describer" [11]. The Describer's goal is to help the Seekers succeed, while the Seekers compete with one another to find the target image within a fixed time limit, using search engine that has indexed the images based on tagging results from the ESP Game. In order to discourage a shotgun approach, the game penalizes Seekers for wrong guesses.

This game goes quite far in capturing the essence of interactive information retrieval. If we put aside the competition among the Seekers, then we see that an individual Seeker, aided by the human Describer and the algorithmic--but human indexed--search engine--is pursuing an information retrieval task. Moreover, the Seeker is incented to be both effective and efficient.

How can we leverage this framework for information retrieval evaluation? Even though the game envisions both Describers and Seekers to be human beings, there is no reason we cannot allow computers to play too--in either or both roles. Granted, the game, as currently designed, focuses on image retrieval without giving the human players direct access to the image tags, but we could imagine a framework that is more amenable to machine participation, e.g., providing a machine player with a set of tags derived from those in the index when that player is presented with an image. Alternatively, there may be a domain more suited than image retrieval to incorporating computer players.

The main appeal of the game framework is that it allows all participants to be judged based on an objective criterion that reflects the effectiveness and efficiency of the interactive information retrieval process. A good Describer should, on average, outscore a bad Describer over the long term; likewise, a good Seeker should outscore a bad one. We can even vary the search engine available to Seekers, in order to compare competing search engine algorithms or interfaces.

7. CONCLUSION

Our goal is ambitious: we aspire towards an evaluation framework that satisfies information scientists as relevant to real-world information seeking, but nonetheless offers the practicality of the Cranfield paradigm that dominates information retrieval. The near

absence of collaboration between the information science and information retrieval communities has been a greatly missed opportunity not only for both researcher communities but also for the rest of the world who could benefit from practical advances in our understanding of information seeking. We hope that the approach we propose takes at least a small step towards resolving this battle royale.

8. REFERENCES

[1] Belkin, N. J., 2008. Some(What) Grand Challenges for Information Retrieval. *ACM SIGIR Forum 42, 1* (June 2008), 47-54.

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

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

[4] Ingwersen, P. and Järvelin, K. 2005. *The turn. Integration of information seeking and retrieval in context*. Springer.

[5] Voorhees, E. 2006. Building Test Collections for Adaptive Information Retrieval: What to Abstract for What cost? In *First International Workshop on Adaptive Information Retrieval (AIR)*.

[6] Turpin, A. and Scholer, F. 2006. User performance versus precision measures for simple search tasks. In *Proceedings of the 29th Annual international ACM SIGIR Conference on Research and Development in information Retrieval*, 11-18.

[7] Saracevic, T. (2007). Relevance: A review of the literature and a framework for thinking on the notion in information science. Part II: nature and manifestations of relevance. *Journal of the American Society for Information Science and Technology 58(3)*, 1915-1933.

[8] Cutting, D., Karger, D., Pedersen, J., and Tukey, J. 1992. Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections. In *Proceedings of the 15th Annual ACM SIGIR International Conference on Research and Development in Information Retrieval*, 318-329.

[9] Workshop on Faceted Search. 2006. In *Proceedings of the 29th Annual ACM SIGIR International Conference on Research and Development in Information Retrieval*.

[10] Von Ahn, L. 2006. Games with a Purpose. *IEEE Computer 39, 6* (June 2006), 92-94.

[11] Von Ahn, L., Ginosar, S., Kedia, M., Liu, R., and Blum, M. 2006. Improving accessibility of the web with a computer game. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 79-82.