HUMAN COMPUTATION FOR HCIR EVALUATION

Shiry Ginosar

Endeca Technologies, Inc.
101 Main Street, Cambridge, MA 02142
sginosar@endeca.com

ABSTRACT

A novel method for the evaluation of Interactive IR systems is presented. It is based on Human Computation, the engagement of people in helping computers solve hard problems. The Phetch image-describing game is proposed as a paradigmatic example for the novel method. Research challenges for the new approach are outlined.

Index Terms— Interactive IR and HCIR evaluation, Web-based games.

1. INTRODUCTION

There are currently two main approaches to evaluation of IR systems - the TREC conference approach and the HCI approach, and neither is optimal across the wide range of systems that exist today. In particular, evaluation paradigms for Interactive IR systems are interesting to investigate, since on the one hand the TREC evaluation method cannot be applied here [7,8] while on the other hand HCI methods tend to be hard to generalize.

In this paper, we consider the relative value of the two primary approaches to this problem and propose and discuss a novel approach to evaluating IR and Interactive IR systems that uses Human Computation [1]. This approach extends TREC evaluation metrics so that it can be applicable to interactive systems, and it improves upon HCI methods by reducing their subjectivity.

2. EXISTING IR EVALUATION METHODS

The first approach for IR systems evaluation, taken by TREC [http://trec.nist.gov] is based on a batch evaluation. The queries and corpus to be used are decided upon *a priori* and the entire corpus is relevance-ranked by hand for each of the queries. Each IR system is then queried using a batch process with the pre-compiled queries over the given corpus. The resulting relevance-ranked set of documents is then compared to the pre-annotated "gold standard" and scores such as *precision* and *recall* are computed [10,13]. The batch process approach is arguably a successful measure of goodness for the effectiveness of the IR system itself [10,11].

However, evaluating an IR system using a batch process may fail to capture the intended use of systems that are designed to support other information discovery processes [13]. This is especially true in regards to evaluating Interactive IR systems. On the one hand, classic IR evaluation relies on a one click paradigm where queries are first composed in full and then sent to the systems to compute a static set of answers [10,13]. On the other hand, Interactive IR systems are often designed to enable a user to iteratively formalize the query. Since the query as a whole is not known *a priori*, there is no way to assess the relevance of documents in the corpus in advance and therefore there is no way to compose a gold standard with which to compare results returned from different systems. Thus, alternative methods must be used in order to evaluate such systems [7, 8].

The second approach to evaluation of IR systems, used primarily within the HCI community, focuses on task level evaluations rather than evaluating the results for individual queries. Such evaluations often employ a mix of objective and subjective metrics such as completion time, user satisfaction and perceived user success [8,14]. Since the metrics used by HCI are partially subjective and since the tasks performed during the evaluation are highly correlated with the specific system and the specific corpus used [8,14], it is hard to compare different systems and the results of these evaluations are rarely accepted by the greater IR community.

Furthermore, HCI evaluations that are set up as user studies are often stymied by the lack of willing participants, the need to compensate participants and the difficulties of recruiting participants from outside the specific university or company where the study is conducted. These hardships can result in lack of data or lack of a sufficiently varied participant population, both of which make it difficult to make statistically significant claims.

3. HUMAN COMPUTATION EVALUATION OF IR

In this paper, we propose a new approach to evaluation of IR and Interactive IR systems. The goal of this line of

thought is to design a system that will allow users to perform search tasks in a natural way, while assessing the quality of the system as well as the success and satisfaction of the users in the background. This approach is not intended to achieve a mapping to the classical evaluation scores used by TREC (unlike [11] which claim to successfully do so, or [7,13] which claim that there is no correlation between user success and TREC metrics). Rather, we seek a new scoring system that will be able to compare different user-systems combinations.

Moreover, to overcome the hardships of recruiting individuals for participation in user studies, we propose to incorporate the concept of Human Computation [1] into the design of our system. Human Computation engages people to aid computers in completing tasks which are either too hard or too expensive for computers to do on their own. Most Human Computation systems are designed as games [1,2,3,4,5,6] which people enjoy playing, or as verification systems which act as gateways to information that people want to access [http://www.captcha.net/, http://recaptcha.net/]. However, a Human Computation system is more than a game: it is cleverly designed such that as a side effect of game play or everyday tasks such as logging in to an email account, useful information can be collected.

4. AN EXAMPLE EVALUATION USING A GAME

As an example of the Human Computation evaluation paradigm, we will investigate in more detail the possible use of the online game Phetch [4] that can be hooked up to different IR systems [3]. Phetch requires players to perform search tasks in order to advance in the game. In Phetch, a describer generates a text description of an image and multiple seekers race to identify the described image out of a large collection of similar images. People play the game because it is fun, and as a side effect of game play the set of IR systems supporting the game may be evaluated. Since the game is interactive in nature, this type of evaluation is suited for IR as well as Interactive IR systems.

There are several advantages for using a game like Phetch for evaluation. First and foremost, since the game involves users performing search tasks while trying to fulfill an information need, it naturally lends itself to evaluation of not only IR systems but also of Interactive IR systems.

Second, the search task itself within the game is done in a natural way. Players are presented with an item (in this case, an image) that they need to find, and they are expected to devise their own ways in which to find it. This type of search task is very similar to search tasks that users of IR systems perform in real life scenarios and therefore would eliminate the need to come up with a contrived simulated work task situation for the purpose of the evaluation [9].

Third, a game like Phetch outputs a clean scoring number for players in the game. This score encapsulates the success of the player both as a seeker who searches for images as well as a describer who describes images for others to find. It depends on the randomly chosen image as the goal of the search, on the speed in which the player processes visual and language information, on the opponents she played against and even on the speed of her internet connection. However, an average scoring over many players, many images and many game sessions could potentially serve as a form of measure of goodness for the combination of a generic user with the specific IR system that was hooked onto the game. This scoring could later be incorporated with other metrics from HCI user studies or batch processes performed against all or part of the IR system to produce a more accurate metric. Repeating the same setting of game play with the same corpus using other IR systems would produce similar scoring which could then be compared with the first, resulting in an overall comparison between the two IR systems and the ways in which they allow users to interact with them.

In this way, a Human Computation system could potentially bridge between the two different approaches to IR evaluation. It could provide a clean score to aid the current ways of comparing different IR systems while also taking into account user interaction with the system as well as the performance of the system itself.

From our experience with Phetch we learned that it can be employed as a possible Human Computation evaluation tool, but an interesting problem is how to apply the concepts from Phetch to non-image domains, in particular text documents.

5. CHALLENGES AND OPEN QUESTIONS

Using Human Computation for evaluation of IR systems requires further research. In particular, this paradigm should be correlated with accepted figures of merit of IR systems that are used by TREC and HCI methods, such as accuracy, precision, recall, success and satisfaction of users.

Additional work may be required for interfacing a Human Computation game with other types of IR systems. For example, in systems that support faceted metadata browsing such as Flamenco [14] and Endeca's [www.endeca.com] Guided Navigation, the corpus should be pre-processed to organize flat tags hierarchically, for which many automatic and semi-automatic methods are available [12]. It is clear that when applying different IR interfaces to the same corpus, the quality of data preprocessing to tailor it to the specific interface could dramatically impact the results of the evaluation. Dealing with preprocessing could potentially be achieved in a manner similar to the one taken by TREC, where competing teams are required to submit a system that interfaces with the Game. The corpus would be known ahead of time, so each team could do their best effort on data preprocessing. Assuming no *a priori* knowledge of the

corpus may also lead to interesting results, but is not currently under consideration.

6. ACKNOWLEDGEMENTS

I thank Daniel Tunkelang, Michael Tucker, Robin Stewart and Andrew Schlaikjer for their insightful comments. I further thank Dr. Luis von Ahn for introducing me to the exciting domain of Human Computation.

7. REFERENCES

[1] von Ahn**,** L. Games With A Purpose. In *IEEE Computer Magazine,* June 2006*,* pp. 96-98.

[2] von Ahn, L., and Dabbish, L. Labeling Images with a Computer Game. In *ACM Conference on Human Factors in Computing Systems (CHI),* 2004, pp. 319-326.

[3] von Ahn, L., Ginosar, S., Kedia, M., and Blum, M. Improving Image Search with Phetch. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* 2007, Vol. 4 pp. IV-1209-IV-1212.

[4] von Ahn, L., Ginosar, S., Kedia, M., Liu, R., and Blum, M. Improving Accessibility of the Web with a Computer Game. In *ACM Conference on Human Factors in Computing Systems (CHI),* 2006, pp. 79-82.

[5] von Ahn, L., Kedia, M., Liu, R., and Blum, M. Verbosity: a game for collecting common-sense facts. In *ACM Conference on Human Factors in Computing Systems (CHI),* 2006, pp. 75 – 78.

[6] von Ahn, L., Liu, R., and Blum, M. Peekaboom: a game for locating objects in images. In *ACM Conference on Human Factors in Computing Systems (CHI),* 2006, pp. 55-64.

[7] Al-Maskari., A. Beyond Classical Measures: How to Evaluate the Effectiveness of Interactive Information Retrieval System?. In *ACM Special Interest Group on Information Retrieval (SIGIR),* 2007, pp. 915.

[8] Borlund, P., The IIR Evaluation Model: A Framework for Evaluation of Interactive Information Retrieval Systems. *Information Research,* 2003, Vol. 8, No. 3.

[9] Borlund, P., and Ingwersen, P. The Development of a Method for the Evaluation of Interactive Information Retrieval Systems. *Journal of Documentation,* 53(3), 1997, pp. 225-250.

[10] Cleverdon, C. The Cranfield Tests on Index Language Devices. *Aslib Proceedings,* 19:173-192, 1967. (Reprinted in K. Spark Jones and P. Willet, editors. *Readings in Information Retrieval.* Morgan Kaufmann Publishers Inc., 1997)

[11] Huffman, S., and Hochster, M. How Well does Result Relevance Predict Session Satisfaction?. *ACM Special Interest Group on Information Retrieval (SIGIR),* 2007. pp. 567-573.

[12] Stoica, E. and Hearst, M. Nearly Automated Metadata Hierarchy Creation. *HLT-NAACL,* 2004. Companion Volume.

[13] Turpin, A., and Scholer, F. User Performance versus Precision Measures for Simple Search Tasks. *ACM Special Interest Group on Information Retrieval (SIGIR),* 2006. pp. 11-18.

[14] Yee, K.P., Swearingen, K., Li, K. and Hearst, M. Faceted Metadata for Image Search and Browsing. In *ACM Conference on Human Factors in Computing Systems (CHI),* 2003.