Evaluating Personal Search Using Games

ABSTRACT

Search in a personal information management (PIM) environment, such as a desktop, is typically evaluated by deploying a system to actual users. These evaluations are expensive and hard to reproduce. As an alternative, in this paper we propose game-style evaluations in which participants complete a set of known-item finding tasks in a competitive environment. Despite limitations, evaluations based on games have numerous benefits, including better experimental control and reusability of the data produced. We describe two game-based user studies we carried out and the lessons learned.

Categories and Subject Descriptors

H.4 [**Database Management**]; D.2.8 [**Information Storage and Retrieval**]: [Information Search and Retrieval]

Keywords

Desktop Search, Evaluation, Human Computation Game

1. INTRODUCTION

It is common for a search system in personal environment to be evaluated by a diary study - deploying the software in a real environment and having it evaluated by actual users [3] [2]. Although this kind of study has its own benefits, it requires lots of resources and the result is often inconclusive in that they relied on subjective judgments as opposed to quantitative analysis. Moreover, the collections and usage logs of these studies are not open to the public, making them hard to be verified by other researchers.

As an alternative way of evaluating personal search, especially known-item finding, we suggest a game-based evaluation method in which participants were asked to find a set of target items in a competitive setting using the system to be evaluated. To simulate a personal search environment, we populate the system with the documents they are familiar with.

SIGIR '10 Workshop on Desktop Search, July 23, 2010, Geneve, Switzerland.

This game-based evaluation method has numerous benefits compared to a traditional user study. First of all, we can induce higher motivation among participants thanks to a more interesting task and the competitive nature of the game. Secondly, good experimental control is ensured since participants are asked to complete a set of given tasks under constraints provided by the game designer. Reusability of data is another benefit considering that public documents are used and most participants are willing to make public their activity logs. Last but not least, developing and running a game-based user study can be done within a relatively small amount of time and efforts, especially when it is implemented on the server side.

One can see that game-based evaluation is not without issues, considering its artificial nature. A competitive environment may lead to a unrealistic behavior. Also, the tasks given to the users are not actual ones in the context of everyday life. Lastly, we are not using personal information that belongs to each user. However, we believe that the advantages outweigh these limitations. Moreover, these can be minimized by sensible design and execution of the study, which we will illustrate using our studies as examples.

The rest of the paper is organized as follows. We briefly explain related work in the next section. Then we introduce the document collection we used and two game-based user studies we performed, focusing on methodology and the data we gathered. Then we describe the lessons we learned through those studies, followed by conclusions.

2. RELATED WORK

To evaluate desktop search, methods for building test collections [1] [4] have been proposed. Among these, the pseudo-desktop method by Kim et al. [4] generated test collections automatically by simulation. Our game-based evaluation technique employs these simulated collections yet improves the procedure of gathering queries and usage logs by providing a game interface which people can enjoy while providing the data needed.

Human computation games [7] have recently been suggested as method for getting a large amount of human annotations in a way that motivates participants using a game-like setting. In the context of IR research, Ma et al. [6] introduced Page Hunt, which is a game designed to collect web search log data by showing each participant webpages and asking her to find them with the search interface provided. Our work essentially adopted their idea of collecting usage log by game interface, yet we made several modifications to simulate search scenarios in a personal environment. Also,

Figure 1: The screenshot of the DocTrack game. The user is being shown a document.

while they used the data for the analysis of usage logs, we used the data for training and testing of retrieval models.

3. OUR METHODS

3.1 CS Collection

Before we delve into the details of user study, we describe the collection we used. We call it a computer science (CS) collection, since the documents of various types are collected from many public sources in the Computer Science department the authors belong to. The CS collection is designed to have the characteristics of desktop data, containing emails from the department mailing list, news articles and blog postings on technology, calendar items of department announcements, webpages and office documents crawled from the department and lab websites.

3.2 DocTrack I

For the first study, our goal was to gather known-item queries in the desktop environment. By adapting PageHunt [6] to our problem setting, we created a game interface called DocTrack as shown in Figure 1. In addition to using documents of many types that might be found in a desktop instead of random webpages, we made several modifications to the original Page Hunt game:

First, since people generally have good knowledge of their own desktops, we collected documents that participants are familiar with and let each of them browse the collection for some time before starting the game; second, to simulate a typical known-item search scenario, we showed participants multiple documents and asked them to find one of them without specifying which one is the target document; third, we used a document viewer that can show documents of any types (e.g. pdf, doc and ppt) in the same way they are seen on the desktop.

Compared to the method of collecting manual queries in Kim et al.[4], using the DocTrack game, we could gather many more realistic queries together with the whole session log data. This in turn allowed us to perform the training of discriminative learning models which typically requires large amounts of training data [5].

3.3 DocTrack II

For our second study, our goal was to evaluate a personal search system where people can use the associative browsing of concepts (e.g., person names, events, etc.) and documents as well as keyword search to find the documents. We modified the DocTrack game used in our first study so that people can browse to related concepts or documents after their initial search.

The study showed that associative browsing plays a role complementary to keyword search during the completion of a known-item finding task. At the same time, we got sufficient amount of data to train and test the learning component of the system.

3.4 Lessons Learned

We learned several lessons during our studies. First of all, we found out that it is important to explain the method of gameplay since users played games online at a convenient location and time, whereas many user studies happen on-site. Especially for our second study, since it involved a new method of finding the target documents, we prepared a screencast explaining the functionality of the game interface in addition to a couple of examples.

We also learned that special attention is required to ensure that playing the game is equivalent to using the system in a natural setting. Although one advantage of game-based evaluation is that users are motivated to get higher scores than other users, this competitive environment sometimes led to behavior that would not usually happen in a natural setting, such as memorizing the whole title of the document. We adjusted the scoring scheme of each search session so that users are discouraged from typing in many keywords.

4. CONCLUSION

In this paper, we suggested a game-based evaluation as an viable alternative to traditional user study. We described its benefits and limitations, and how we could successfully employed such method in two user studies we performed.

5. ACKNOWLEDGEMENTS

This work was supported in part by the Center for Intelligent Information Retrieval and in part by NSF grant #IIS-0707801. Any opinions, findings and conclusions or recommendations expressed in this material are the authors' and do not necessarily reflect those of the sponsor.

6. REFERENCES

[1] S. Chernov, P. Serdyukov, P.-A. Chirita, G. Demartini, and W. Nejdl. Building a desktop search test-bed. In *ECIR*, pages 686–690, 2007.

[2] E. Cutrell, D. Robbins, S. Dumais, and R. Sarin. Fast, flexible filtering with phlat. In *CHI '06: Proceedings of the SIGCHI conference*, pages 261–270, New York, NY, USA, 2006. ACM.

[3] S. Dumais, E. Cutrell, J. Cadiz, G. Jancke, R. Sarin, and D. C. Robbins. Stuff i've seen: a system for personal information retrieval and re-use. In *SIGIR '03*, pages 72–79, New York, NY, USA, 2003. ACM.

[4] J. Kim and W. B. Croft. Retreival experiments using pseudo-desktop collections. in Proceedings of CIKM'2009, Hong Kong, China, 2009.

[5] J. Kim and W. B. Croft. Ranking using multiple document types in desktop search. In *To Appear in SIGIR '10*, New York, NY, USA, 2010. ACM.

[6] H. Ma, R. Chandrasekar, C. Quirk, and A. Gupta. Improving search engines using human computation games. In *CIKM*, pages 275–284, 2009.

[7] L. von Ahn and L. Dabbish. Designing games with a purpose. *Commun. ACM*, 51(8):58–67, 2008.