User-model-based Evaluation for Interactive Image Retrieval

Masashi Inoue* and Manh Hong Nguyen†
**Yamagata University, Yonezawa, and National Institute of Informatics, Tokyo, Japan*
†*Viettel Telecom, Hanoi, Vietnam*

***Abstract*—User-system interaction is sometimes a cumbersome element of non-textual information access. Image retrieval systems now incorporate various interaction mechanisms. However, there is no standard or established way of interacting that has been proven to be most effective. Therefore, many system designers are introducing or eliminating various functionalities. The bottleneck in the development effort is finding an effective and fair way to evaluate these systems. For static information access, several test collections have commonly been used. For interactive information access, expensive laboratory testing is needed. This paper presents a user semi-automatic user-model-based evaluation method that is designed to reduce cost while maintaining objectivity.**

***Keywords*-evaluation; image retrieval; user model;**

I. Introduction

Users of information access systems rarely have a concrete idea of their search needs or the location of relevant information sources. This is especially true when information is in the from of non-textual media such as images because non-textual information retrieval is not as effective as textual retrieval and people are not used to representing their non-textual search needs precisely. As a result, in many search sessions, users cannot complete their task in a single operation; they must undertake several interactions with the system.

The difficulty in developing an interactive system that can accommodate such user behavior lies in the performance evaluation. It should be cost-efficient and fair. In static information access, system performance is usually verified using shared test collections. By using the same queries, documents, and relevance assessments, systems developed by different designers can be evaluated relatively fairly and automatically. Further, there have been attempts to reduce the cost of building collections by using automation [1]. However, such test collection cannot be built for interactive systems and laboratory testing is needed. On one hand, by hiring a sufficient number of participants to take part in the testing, relatively accurate performance can be measured. However, such laboratory evaluation is time-consuming and expensive. On the other hand, having the designer of the system evaluate it can be very efficient. However, such subjective evaluation may be biased, and the system may not be relevant to generic users. Therefore, in this paper, we propose the utilization of user models that are built on the actual user-system interaction history acquired in previously evaluated systems. By doing so, we can simulate the behavior of unseen users of newly developed systems. This user-model-based evaluation can incorporate the individualities of users while reducing the cost. To improve current systems that have already attracted many users, such as widely used web search engines, developers can regard the randomly selected existing users as participants in laboratory testing and evaluate new functionalities or designs. In contrast, if systems are new and do not yet have any users, evaluations must be done by the developers themselves. To alleviate the cost problem associated with actual users, the use of imaginary users called "Personas" has been proposed [2]. Persona have different personalities and backgrounds in terms of age, sex, job, life-style, and preferences, which were created by the designers. They are expected to use the systems and interfaces according to their imaginary personalities. Although the use of personas can reduce the cost of actual user testing, the outcome of the evaluation is qualitative, and the personalities are limited by the designaer's imagination. For more quantitative model-based evaluation, a Markov model to simulate the users action in button-pushing interfaces has been tested [3]. The idea is that the current button-push action depends only on the previous actions and the users select the next action probabilistically. The task is to design the probability distributions over possible actions under the interface design being tested. Similarly, a Markov model is used to visualize user behavior in personal infromation management[4]. We have developed a similar framework to enable user modeling of interactive image retrieval systems.

II. Experiment

A. Image Collection

We used the ImageCLEF2008 ad hoc photo retrieval collection [5]. There were 39 search topics and $40,000$ images, and each image had a textual annotation. Search activities were initiated by using "content" fields of the topics as textual queries. Since this was a test collection, each search topic had a defined relevant image set. Therefore, we were able to quantitatively measure how many relevant images a participant had collected at each iteration of interaction.

B. Participants

Twenty-four participants joined the user testing. They ranged in age from 19 to 32 and were technicians, students, or teachers who had a certain operational knowledge of computers. Participants were asked to collect as many relevant images as possible for given search topics. That is, the goal is to obtain a high recall rate in this search scenario. They could stop searching when they were satisfied with the retrieval results or did not want to continue the search anymore. Therefore, for the same search topic, the number of interaction steps and collected relevant images varied among participants. For the filtering purpose, of the 24 participants, we selected the top 10 who clearly tackled the task seriously. The operation logs of those 10 participants were used as the data for model building.

C. Prototype System

A prototype interactive image retrieval system was developed to collect the user action logs, user information, and internal parameter values of the system. The system can be accessed over the Internet. Firefox was the recommended browser. Each participant who joined our experiment was asked to create a user account before starting a search. Participants selected one of 40 queries defined in the test collection to initiate the search. In the system, users could take one of the following three actions at each iteration. **Find Similar** (S): searches for visually similar images to the selected image. **Query Reformulate** (R): lets users reformulate textual queries for the next search. **Go Back** (B): navigates users to the previous result page. In the S operation, the system finds the similar images based on the color information extracted from images. In the R operation, the user changes a textual query by selecting one of the candidate queries in the phrase list provided on the page.

D. Analysis of collected data

An important element of the collected usage log is the number of interactions with the system. The frequency distribution of the round at which the 10 participants stopped their search over 299 search trials is shown in Figure 1. Note that the total number of trials is not the product of the number of participants and that of tasks. Some participants may execute several trials for a single search topic. Also, some participants did not cover all search topics. As the number of search rounds increases, the frequency decreases monotonically. The maximum number of rounds 25 is observed only once, with the mean occuring at 7.28. The users interacted with the systems a relatively long time because the search scenario was to collect as many relevant images as possible.

E. Model Generation

Based on the usage history data collected, as described above, we built a probabilistic model that explains how likely the next action is given the current user status. For example, if the next action is only dependent on the interaction round r, the model for all users is estimated as follows:

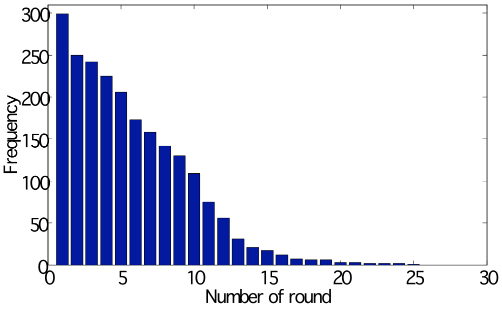

Figure 1. Histogram of the number of interactions by the 10 participants in 299 trials.

$$\begin{aligned} P(S,R,B|r=2) &= (0.339, 0.553, 0.128) \\ P(S,R,B|r=4) &= (0.272, 0.602, 0.126) \\ P(S,R,B|r=6) &= (0.253, 0.652, 0.095) \end{aligned}$$

This indicates that as the search goes on, generic users prefer textual queries over searches using visual similarity.

III. Conclusion

This paper proposed a model-based evaluation framework for use in developing interactive multimedia systems, especially image retrieval systems. The proposal includes a method to collect usage logs from actual users and a method to build a model from the collected logs. The characteristics of the usage logs were also analyzed. The next stage of our research involves generating pseudo-user action sequences from the acquired user models that are usable to evaluate systems.

References

[1] E. Graf and L. Azzopardi, “A methodology for building a patent test collection for prior art search,” in *Proceedings of the Second International Workshop on Evaluating Information Access (EVIA)*, December 2008, pp. 60–71.

[2] A. Cooper, *About Face 3: The Essentials of Interaction Design*. Wiley, 2007.

[3] H. Thimbleby, “User interface design with matrix algebra,” *ACM Trans. Comput.-Hum. Interact.*, vol. 11, pp. 181–236, 2004.

[4] D. Elsweiler, M. Hacker, and S. Mandl, “Visualising pim behaviour with markov chains,” in *Personal Information Management: PIM 2009*, Vancouver, Nov. 2009.

[5] T. Arni, P. Clough, M. Sanderson, and M. Grubinger, “Overview of the ImageCLEFphoto 2008 photographic retrieval task,” in *Evaluating Systems for Multilingual and Multimodal Information Access – 9th Workshop of the Cross-Language Evaluation Forum*, ser. LNCS, vol. 5706, Sep. 2009, pp. 500–511.