An Application of the NTCIR-WEB Raw-data Archive Dataset for User Experiments

Masao Takaku
Research Organization of Information and System
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, Japan
masao@nii.ac.jp

Yuka Egusa
National Institute for Educational Policy Research
6-5-22 Shimomeguro, Meguro-ku, Tokyo, Japan
yuka@nier.go.jp

Hitomi Saito
Aichi University of Education
1 Hirosawa, Igaya-cho, Kariya-shi, Aichi, Japan
hsaito@auecc.aichi-edu.ac.jp

Hitoshi Terai
Nagoya University
Furo-cho, Chikusa-ku, Nagoya, Aichi, Japan
terai@nul.nagoya-u.ac.jp

Abstract

This paper presents a simple approach to utilize past test collections as a material for user experiments. We have built a Web-based user interface for NTCIR-5 WEB run results, and conducted a user experiment with 29 subjects to investigate whether performance evaluation metrics of information retrieval systems used in test collections such as TREC and NTCIR comparable to user performance. In this experiment, we selected three types of systems from among systems that participated in NTCIR-5 WEB, and then selected three topics with roughly the same values from among several search topics. The results of the experiment showed no significant differences among these systems and topics in the time for search. While, in general, the user experiment itself have been successfully conducted and shown similar trends with prior study, the approach seems to have some limitations mainly on interactivity and cached page display.

Keywords: *Evaluation, User experiments, User interface, Rawdata archive, Web information retrieval*

1 Introduction

Performance evaluations for information retrieval (IR) systems are extremely important in today's Internet environment, where a wide variety of IR systems are provided and used. The evaluations of IR systems have begun with Cranfield's experiments and the field later expanded to evaluation experiments using large-scale test collections, such as TREC[1] and NTCIR[2]. And then, these methods have been deployed as a "standard method" on a wide variety of system evaluations on information access technologies including IR, question-answering, summarization and so on.

[1] http://trec.nist.gov/
[2] http://research.nii.ac.jp/ntcir/

In recent years, however, these evaluation methods have been called into question. Several researches [1][3][4] suggest that the results of performance metrics in past system evaluations do not necessarily match the results of subjective evaluations and perception characteristics in user evaluations. There has not been sufficient study, however, into why these results do not match, or what can be done to achieve performance evaluations that are closer to the users' evaluations.

The reasons why these kinds of user experiments have not been investigated are caused by: (1) There is little environment that effectively connect system evaluation metrics with user experiments. (2) User experiments themselves are time-consuming and expensive. (3) Effective experimental design is needed beforehand.

Based on the above situation, in this research, we took an approach to utilize the NTCIR-5 WEB raw-data archive dataset for user experiments. This approach directly combines past results on test collection with user experiments. We constructed a Web-based user interface based on it, aiming at comparing user evaluations with batch evaluations in the NTCIR-5 WEB Navigational Retrieval task (Navi2) [2].

2 Related works

Hersh et al. [1], and Turpin and Hersh [3] reported that in the TREC 7-9 Interactive Track, batch system evaluations did not correspond to the results of user evaluations. In these experiments, they used two phase model for separate system evaluation and user experiments. In first phase, several IR systems were evaluated on batch environment, and then user experiments were made on those systems.

In 2006, Turpin and Scholer [4] reported similar trends with a simple Web search task. In Turpin's work, ranked lists of retrieved documents were auto-

matically created based on the settings of their precision/recall metrics. That introduced user experiments without retrieval engines and iterative large-scale experiments.

3 Approach and user interface

Figure 1. Search result interface for a query

We created a Web-based user interface for NTCIR run results. Figure 1 shows our search result interface. This search result page shows ten documents at one time, and these pages are derived from NTCIR-5 WEB submitted runs. The run results, which were submitted by participants for the task, consist of a ranked document list. We just simply made a list of documents along with each ranked list, and showed its summary, which consists of a rank, title, snippet text, URL, and document ID of a page. Each snippet shows a context summary based on the query keywords. These titles and URLs are linked into the corresponding cached pages.

Cached pages show snapshot of Web pages from NTCIR-5 WEB test collection (NW1000G-04). For cached page, we reused the same display engine as the relevance judgements of NTCIR-5 WEB. All the links in pages are redirected through within the cached page space. If a referring page is not included within NW1000G-04, that link shows a warning message: "There is no corresponding page of this link. Five seconds later, you will be redirected into the present URL." Then it automatically redirect to the present page on the real Web. Note that NW1000G-04 does not gather all the images and several types of document format, such as PDF and MS-Word, other than HTML and plain text. In the case of a site using those images or other navigational elements like Flash format, it might be difficult for users to understand the content and the site structure.

4 User experiments

The user interface described in Section 3 were deployed for our user experiments. In this section, we will describe its experimental design and some results.

4.1 Experimental design

A total of 29 subjects (19 male, 10 female) participated in the experiment. The average Internet usage time for all subjects was 3.01 hours/day ($SD = 2.55$).

The experiment was conducted using a 3x3 mixed design. The first factor was the three topics, and the second factor was the three systems. As indicated in Table 1, the subjects were allocated into three patterns (S_a, S_b, S_c) combining search topics (movie, shopping, restaurant) and systems (high, middle, low). During the experiments, the subjects were randomly assigned for each pattern, and each pattern had nine or ten subjects.

Table 1. Experimental design

	High	Middle	Low
Movie	S_a	S_c	S_b
Shopping	S_b	S_a	S_c
Restaurant	S_c	S_b	S_a

4.2 Materials

Three topics and three systems were selected from the NTCIR-5 WEB task for use in this experiment.

From among the systems participating in the NTCIR-5 WEB task, three systems were selected as having normalized discounted cumulative gain (nDCG) values and reciprocal-rank (RR) values corresponding to High, Middle, and Low (TNT-3, ORGREF-C20-P2, and ORGREF-GC1, respectively). Three topics (movie, shopping, and restaurant) were selected as having similar nDCG values within a single system (topic numbers 1196, 1296, and 1367, respectively).

Figure 2 shows the systems' nDCG and RR values for each topic. High had high nDCG and RR values for all topics, and Low had low nDCG and RR values for all topics. During the each search task, the subjects were not informed that they were using different search systems each time.

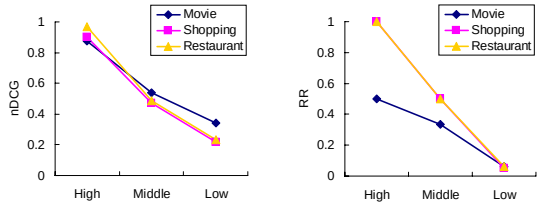

Figure 2. nDCG and RR values for each system and each topic

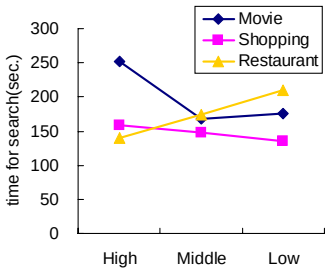

Figure 3. Average search time for each system and each topic

4.3 Procedures

As described in Section 3, we used raw data archive dataset for the user experiments. In this section, details of the user experiments will be described.

First, the subjects were given a questionnaire on their experience in using the Internet and computers. After an introduction to the search tasks, subjects performed a practice search. After this, the topics were presented in random order according to the conditions of the experiment shown in Table 1. The search topics were displayed on a Web browser. When the search began, the following information was displayed: The purpose of the search, Background, Required conditions and the link to the search result pages described in Section 3. The subjects could jump to the search result pages whenever they wanted. From the search result interface, the subjects looked for pages that appeared to match the topic context. The search ended when the relevant page was found, and subjects were asked to evaluate the search.

4.4 Results

Figure 3 shows the subjects' average search time in seconds for each system and each topic. From this plot, we can see that for Movie and Shopping, High had the longest execution time, but in the case of Restaurant, the search time grew longer from High to Low. There was no significant difference, however, between systems and topics.

As in the case of prior research, these results suggest that even when evaluation data in the NTCIR-5 WEB task is used, the results of system performance based on batch evaluations do not match the results of user performance in user experiments.

5 Discussion

Our approach described in Section 3 was used in the user experiments. Our user experiments using the model was successfully conducted, and showed similar trends with prior research [3]. That implies the usefulness of the model itself.

Our approach utilizes the raw dataset of the run result from the participants at NTCIR-5 WEB. The same approach can be applied to other test collections, including NTCIR, TREC and CLEF and so on. And therefore these efforts could be connected with user experiments by using our approach.

In general, user experiments for system evaluation needs several types of efficient and effective retrieval systems. For clarifying experimental design, system evaluation usually needs a "baseline" (controled) and "improved systems" (experimental), in terms of experimental criteria. Comparing relative performance in interactive systems needs some level of efficiency in system response time. It would be difficult, however, to build such efficient systems, because today's test collection has a huge number of documents (e.g. NW1000G-04 has approximately 100 million web pages; 1.36TB [2]). Additionally, in experiments, the results can be biased if a small number of similar retrieval engines are used. So, comparing different retrieval schemes needs quite different retrieval engines. That will introduce another type of difficulties on building and testing several retrieval engines. Our simple approach, therefore, could remove costs to prepare those retrieval engines.

Other approach includes interactive search engines which takes user's queries online. That approach has advantages than ours in terms of realistic search environment, but it also needs to take costs to make efficient/effective retrieval systems and to make additional optional relevance judgements. And also if a users experiment allows arbitrarily user-entered queries, it will happen to appear new unjudged documents in search result pages. For a stable evaluation, those unjudged documents need additional relevance judgements. Turpin and Scholer [4] took another approach to build search results automatically from pooling document lists based on precision/recall measures in order to avoid this.

After the experiments, We asked subjects several questions about impressions of the experiments. For example, some subjects identified the experimental

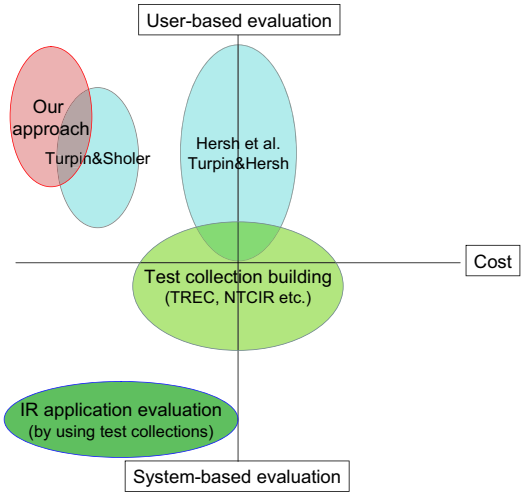

Figure 4. Comparison of environment for user experiments by considering costs and evaluation preferences

environment different from usual search engines: "I didn't have an impression that I was doing a search by myself."; "It was difficult to search because only search result pages are available."; "I felt really uncomfortable with these experiments of searching a relevant page only from the search result pages of the same keyword." And some complained about cached page schemes: "In the experiments, there were many pages that redirected to the not-found-pages, and pages that did not display images correctly. So, it was difficult to judge the page relevant or non-relevant."; "In a topic, since some images did not appear, that made the task difficult to do."

As indicated in these comments, our approach has limitations mainly for non-interactive design and cache-based experiments. For experimental environment, interactive retrieval experiments will be the key to investigate more real situations. And for issues on cached pages, one approach will be crawling all the page components including images and other object data files for more real Web experimental settings. Another approach will be to use the real Web directly.

While our approach utilizing the raw-data archive was successfully applied to the user experiments, we will further investigate experimental design and ways to utilize a raw-data archive dataset.

Figure 4 shows a rough comparison of our approach with prior studies in terms of costs to run user experiments, and user-based or system-based evaluation. Existing test collection building projects and its deployment into IR application research are common. But few works on combining those test collections with user-based evaluation were made. Our approach is simple, but brings user-based IR evaluations into the existing test collection.

6 Conclusion

We took an approach of utilizing raw-data archive dataset in a past test collection workshop, in order to facilitate user experiments for system evaluation. We built a Web-based user interface for user experiments based on the approach. Then, a user experiment was successfully applied based on the interface.

From the results of our experiment, in the case of NTCIR-5 WEB Navigational Retrieval task, nDCG/MRR system performance measures did not match with users' performance evaluations. These results seem to be similar with prior works, and suggest our approach could be a reasonable option on user experiments.

Although the approach itself has a limitation on an interactive data from user experiments, this simple approach can be easily taken without effective retrieval systems and additional relevance judgements. This approach facilitates not only utilization of existing test collection datasets but also user experiments.

In the future, we need more analysis on experiments based on this model, such as analysis on user tracking logs and so on.

Acknowledgment

This research was supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Scientific Research of Young Scientists (B), No.17700604; No.17700130.

References

[1] W. Hersh, A. Turpin, S. Price, D. Kraemer, D. Olson, B. Chan, and L. Sacherek. Challenging conventional assumptions of automated information retrieval with real users: Boolean searching and batch retrieval evaluations. *Information Processing & Management*, 37(3):383–402, 2001.

[2] K. Oyama, M. Takaku, H. Ishikawa, A. Aizawa, and H. Yamana. Overview of the NTCIR-5 WEB navigational retrieval subtask 2 (Navi-2). In *Proceedings of NTCIR-5 Workshop Meeting*, pages 423–442, 2005.

[3] A. Turpin and W. Hersh. Why batch and user evaluations do not give the same results. In *SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 225–231, 2001.

[4] A. Turpin and F. Scholer. User performance versus precision measures for simple search tasks. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 11–18, 2006.