Comparing System Evaluation with User Experiments for Japanese Web Navigational Retrieval

Masao Takaku
Research Organization of Information and Systems
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

We conducted a search experiment targeting 31 users to investigate whether the performance evaluation metrics of IR systems used in test collections, such as TREC and NTCIR, are comparable to the user performance and subjective evaluation. We selected three systems with high, medium, and low performance values in terms of nDCG, MRR and Prec@10 metrics from among the retrieval systems that participated in the NTCIR-5 WEB task, and then selected three topics. The results of the experiment showed no significant differences between these systems and topics in the completion time for each search. Furthermore, none of the results of the users' evaluations corresponded to the results of the batch system evaluations. These results indicate a need for new evaluation metrics that correspond to the users' evaluations.

1. INTRODUCTION

The 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 performance evaluations of IR systems are said to have begun with the Cranfield experiments, and the field later expanded to include evaluation experiments the use large-scale test collections, such as TREC and NTCIR.

In recent years, however, these evaluation methods have been called into question. In researches conducted by Hersh et al.[1], and Turpin and Hersh[4], it was reported that in the TREC 7-9 Interactive Track, batch evaluations did not correspond to the user evaluation results. Turpin and Shoeler[5] recently conducted more large scale tests on a simple Web information finding task, and showed that the system's MAP metrics and user performance did not correlate with each other.

This results 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. However, there has been little study that has focused on this gap between the batch and user evaluations. It is necessary to gather evidence using other types of tasks or test collections to investigate why batch and user evaluation do not match, or what can be done to develop performance evaluations that are closer to the users' evaluations. Most of the previous researches have been based on TREC data, and there have been almost no studies using other large-scale test collections.

Based on the above situation, we compared user evaluations with batch evaluations in the NTCIR-5 WEB Navigational Retrieval task (Navi2)[2] for our current research. We report the preliminary results from our experiments in this paper, and introduce about the differences and similarities between our results and those of prior researches.

WISI 2007 XX XXXX, XXXXX

2. METHODS

2.1 Subjects and Design

A total of 31 subjects (21 males and 10 females) participated in the experiment. The subjects were recruited from three universities; 12 were faculty members, 8 were graduate students, and 11 were undergraduate students. The backgrounds of the subjects varied, but the faculty members were from the nursing science field, the graduate students were from the information science field, and the undergraduate students were from the education science and information science fields. The average age of the subjects was 25.6 ($SD = 4.99$), and the average Internet usage time was 2.98 hours per day ($SD = 2.43$). They were unfamiliar with our dataset and this was their first time to use it.

The experiment was conducted using a 3×3 mixed design. The first factor was the three topics, and the second factor was the three systems (both were subject internal factors). As indicated in Table 1, the subjects were allocated into

three patterns (S_a, S_b, and S_c) combining the topics (movie, shopping, and restaurant) and systems (high, middle, and low). During the experiments, the subjects were randomly assigned to each pattern, and each pattern had ten or eleven 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

2.2 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), reciprocal-rank (RR), and precision at 10 (Prec@10) 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[1]). Figure 1 shows an English translation for the shopping topic.

```
<TOPIC><NUM>1296</NUM>
<TITLE>Seiyu, online supermarket</TITLE>
<DESC>I want to visit to Seiyu's online supermar-
ket page.</DESC>
<NARR>
<BACK>I would like to go to shopping at Seiyu's
online supermarket.</BACK>
<RELE>Seiyu's online supermarket page in the offi-
cial Seiyu website is relevant.</RELE>
</NARR>
</TOPIC>
```

Figure 1: Shopping topic (English translation)

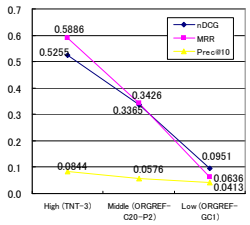

Figure 2: nDCG, MRR and Prec@10 performance measures of three runs with 269 topics

Figure 2 shows the systems' nDCG, MRR and Prec@10 values for all 269 topics from the NTCIR-5 WEB, and Fig. 3 shows these measures for each selected topic. In the NTCIR-5 WEB, graded relevance levels are assigned in relevance judgments. We calculated these evaluation metrics based on a weighted values. For nDCG, multiple relevant levels were weighted with (A, 10) and (B, 1). For MRR and Prec@10, multiple relevant values were calculated at rigid level: (A, 1), (B, 0). Additionally, duplicates of the relevant documents were judged at relevance judgments. For nDCG and Prec@10, if duplicate relevant documents were found several times, they were regarded as irrelevant except the first one found. In terms of the NTCIR-5 WEB test collection having 269 topics, as shown in the Fig. 2, there were significant differences between these three runs (high, middle, and low), and these results were confirmed by using a pair-wise t-test for the three evaluation metrics (all results at $p < .001$).

[1]All topics are available at the NTCIR website: http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings5/cdrom/WEB/NAVI2/ntcweb5-navi-frun-topics-1.euc.txt (only in Japanese)

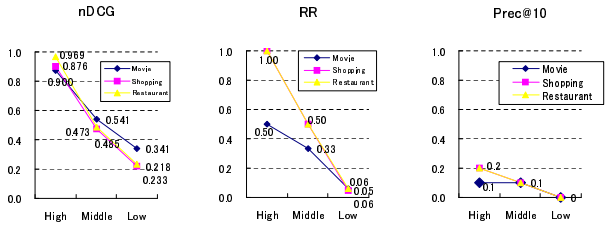

Figure 3: nDCG (left), RR (middle) and Prec@10 (right) performance measures of three runs for each topic

As can be seen in Fig. 3, the high run had the highest nDCG, MRR and Prec@10 values among the three topics, and the low run had the lowest nDCG and RR values among the three topics.

2.3 Procedures

Figure 4: Search result interface for a query

During the search experiment, the subjects were instructed to read a topic's description, background, and relevant criteria, and then to explore the Web, which is in reality the NW1000G-04 dataset, through our Web-based user interface for NTCIR-5 WEB run results[3] to find a relevant page. We instructed the subjects to bookmark the relevant page if they found a relevant page, and then their task for the topic would be complete. Our Web-based interface is shown in Fig. 4. We assumed that these settings could partially sim-

ulate their daily search environment in a real Web search engine.

First, the subjects were given a questionnaire on their demographics and experiences in using the Internet and computers. After an introduction to the search tasks, the 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. During each search task, the subjects were not informed that they were using different search systems each time, because we tried to assure the subjects not to have a bias against the systems, and to keep their mind neutral during the experiment. The search topics were displayed on a Web browser. When the search began, the following information was displayed: The purpose of the search (<DESC> in Fig.1), background (<BACK> in Fig.1), relevance criteria (<RELE> in Fig.1) and the link to the search result pages (SRPs). The subjects could jump to the SRPs whenever they wanted. The SRP was composed of a list of ten pages at one time, and its interface was similar to that of a usual search engine which has a title, URL, and snippets of pages (See Fig. 4). Note that we created and used the static SRPs from the search result runs submitted to NTCIR-5 WEB. The subjects looked for pages that appeared to match the topic context from this list of SRPs. The search ended when the relevant page was found, and the subjects were asked to evaluate the search. The searches for each topic were evaluated using a 5-point scale based on the following items: (1) Search difficulty, (2) Satisfaction with the results, (3) Confidence in the results, (4) Appropriateness of the system for that topic, and (5) Prior knowledge of that topic.

At the end of the experiment, the subjects were informed that a different search system had been used for each of the three topics, and they were then asked to fill out the following two evaluations using a 3-point scale: (1) Performance of the three systems, and (2) How difficult it was to understand the search itself.

3. RESULTS AND DISCUSSIONS

3.1 Agreement with the official assessments

In general, the relevance judgments between people presented quite differing results[6]. From our experience with Web navigational retrieval, the agreements between people are rather high for the navigational task than for the other tasks (e.g. informational task).

Table 2: Agreement rates at rigid (top) and relaxed (bottom) level between subjects' judgments and official judgments

Rigid level:

	High	Middle	Low	(total)
Movie	3/10	2/10	0/11	5/31
Shopping	7/11	8/10	8/10	23/31
Restaurant	10/10	10/11	6/10	26/31
(total)	20/31	20/31	14/31	54/93

Relaxed level:

	High	Middle	Low	(total)
Movie	9/10	8/10	11/11	28/31
Shopping	8/11	8/10	9/10	25/31
Restaurant	10/10	10/11	6/10	26/31
(total)	27/31	26/31	26/31	79/93

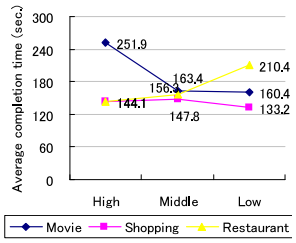

Figure 5: Average search time for each system and topic

In our experiments, the subjects reported on a relevant page. The agreement rates between the relevant pages reported by the subjects and the official assessments from the NTCIR-5 WEB[2] are shown in Table 2. The differences among the systems were tested by using the Chi-square test and two-way ANOVA, but no significant differences were observed. Only at the rigid level[2] was there a significant main effect of the topic ($F(2, 84) = 26.887, p < 0.001$), and the Movie topic showed significantly lower agreement rates than those of the official relevance judgments ($MSe = 0.152, p < 0.001$). At the rigid level, the agreement rates on the topic Movie were rather low, although the agreements on the other topics were quite high. At the relaxed level, on the other hand, the agreement rates were high for all the topics. The low agreement rate for the Movie topic was caused by its absence in the relevance criteria in its topic description. However, most of the subjects could find at least one relevant or partially relevant page. From the relaxed level results, our results seem to be reasonably consistent with those from the original NTCIR-5 WEB, and we could see that our experiment settings could successfully simulate the navigational retrieval settings of the original NTCIR-5 WEB.

3.2 Completion time

Figure 5 shows the subjects' average search completion time in seconds for each system and topic. We can see from this plot that for the movie and shopping topics, the high run had the longest execution time, but in the case of the restaurant topic, the search time grew longer from the high to the low run. There was no significant difference, however, between the systems and topics.

These results suggest that even when the evaluation data in the NTCIR-5 WEB task is used, the system performance results based on the batch evaluations do not match the results of the user performance in the user experiments. The search completion time is one of the user performance measures, because one of the purposes of information access systems for end users is, in general, to quickly retrieve information. From this viewpoint, our results for the search completion time show that subjects can get information almost in the same period of time whether or not they use a batch high-performance system.

Turpin and Hersh[4] reported that users performed equally well on significantly different batch evaluation systems in

[2]The NTCIR-5 WEB had graded relevance judgments, in which a document was assessed as relevant, partially relevant, or irrelevant. In a rigid level analysis, partially relevant documents are seen as irrelevant documents, and in the relaxed level, they are regarded as relevant.

Table 3: Summary of subjective evaluation analysis by two-way ANOVA

	System×Topic	System	Topic		Main effect			MSe	p-value
	p-value	p-value	p-value	F(2,84)					
Difficulty	0.276	0.364	0.057+	2.972	Restaurant	>	Shopping	1.238	0.023*
Satisfaction	0.798	0.210	0.052+	3.703	Shopping	>	Movie	1.019	0.015*
Confidence	0.742	0.771	0.021*	4.058	Shopping	>	Movie	0.961	0.001**
Appropriateness	0.559	0.934	0.023*	3.961	Shopping	>	Restaurant	1.127	0.001**
Prior knowledge	0.353	0.525	0.183	1.734	—		—	—	—
Performance	0.733	0.730	0.000**	8.894	Movie	>	Restaurant	0.481	0.000**
					Shopping	>	Restaurant		0.000**
Difficulty in understanding	0.843	0.105	0.013*	4.577	Movie	>	Shopping	0.523	0.003**

+: $p < 0.1$, *: $p < 0.05$, **: $p < 0.01$

terms of MAP. Turpin and Shoeler[5] also reported that the MAP values did not match with their user performances, based on the automatically created rankings. The differences between those studies and the current one are threefold. First was the dataset that was used. We used the NTCIR-5 WEB dataset, which consists of Japanese topics and mainly a Japanese Web dataset, while prior researches used a TREC dataset, which consists of English topics and newspaper articles or a Web dataset in English. Second was the tasks that were conducted. We conducted experiments for a Web navigational retrieval task, while prior researches were either for a recall-oriented task, a Q&A task, or a Web information-finding task. The third difference was in the evaluation metrics that were used. The NTCIR WEB used a multi-graded relevance level, and reported the DCG and MRR as its official metrics. We used the nDCG, MRR, and Prec@10 as the system evaluation metrics, while prior studies used the MAP and Prec@n as major metrics for the batch evaluation.

Although the datasets, tasks, and metrics were changed, their results and ours are quite similar. That is, from the user experiments, we found that the difference in batch system evaluations that were used does not directly result in the differences in user performance, which is measured by the time taken to complete each task.

3.3 Subjective evaluation

We conducted two-way ANOVAs with topics and systems as the between-subject factors regarding the seven subjective evaluation points answered by the subjects (task difficulty, satisfaction with the result, confidence in the result, appropriateness of the system, prior knowledge of the topic, system performance, and difficulty in understanding of the task).

The results from a statistical test for the systems and topics are given in Table 3. The results from this analysis showed that several significant differences were found between the topics, except for the prior knowledge of a topic. However, it did prove that no significant differences were found among the systems, or among the topics×systems. In summary, these results suggest that users are more aware of the differences between the topics than they are of the differences in the performances of the different systems.

From the subjects' comments, we noticed that for some topics there was difficulty in finding a relevant document. In some parts, this was caused by the limitation in the experimental environment. For example, the restaurant topic was more difficult than the other topics in terms of difficulty, appropriateness, and performance. The relevant pages for the restaurant topic were easily found and the relevant pages had several embedded images in them, but the NW1000G-04 dataset only gathered in text format. So, the subjects noted a poorer performance for the restaurant topic. Another example is with the movie topic. The movie topic does not have relevance criteria in its topic description. So, the subjects noted lower scores in the confidence, satisfaction, and ease in understanding for the movie topic.

4. CONCLUSION

In looking at our experimental results, in the case of the NTCIR-5 WEB task, the nDCG, MRR, and Prec@10 system performance measures did not match the users' performance and subjective evaluations. These results could be viewed as suggesting a need for the development of new evaluation metrics that more closely correspond to the user evaluations. In the future, we will analyze the subjects' tracking log data during the experiments and other supporting information. In addition, since the size of the topics we used was small, our analysis could be made more stable if we had more topics. We will test this point in the future.

5. ACKNOWLEDGMENTS

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

6. REFERENCES

[1] W. Hersh et al. 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 et al. Overview of the NTCIR-5 WEB navigational retrieval subtask 2 (Navi-2). In *Proceedings of NTCIR-5 Workshop Meeting*, pages 423–442, 2005.

[3] M. Takaku et al. An application of the NTCIR-WEB raw-data archive dataset for user experiments. In *Proceedings of EVIA 2007 (NTCIR-6 Pre-Meeting Workshop)*, pages 78–81, Tokyo, 2007.

[4] A. Turpin and W. Hersh. Why batch and user evaluations do not give the same results. In *Proceedings of SIGIR 2001*, pages 225–231, 2001.

[5] A. Turpin and F. Scholer. User performance versus precision measures for simple search tasks. In *Proceedings of SIGIR 2006*, pages 11–18, 2006.

[6] E. M. Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. In *Proceedings of SIGIR 1998*, pages 315–323, 1998.