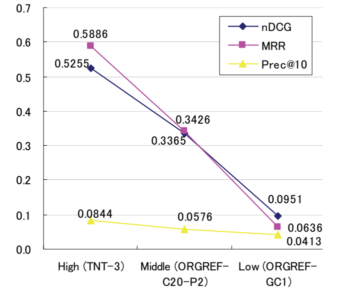
An Application of the NTCIR-WEB Raw-data Archive Dataset for User Experiments

Purpose of our work

- Confirm prior researches[1][2] for evaluation metrics and user performance.
- Get some insights in evaluation metrics and user performance with Terabyte-scale Web collection (NTCIR-5 WEB[3][4]).
- Utilize the NTCIR-5 WEB Raw-data Archive dataset for user experiments.

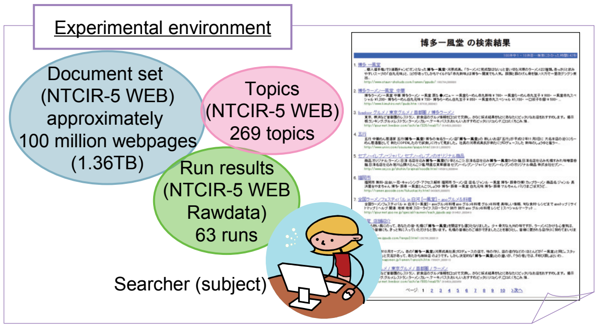
User Experiment

- 31 subjects (21 male, 10 female) were recruited from three universities; 12 faculties, 8 graduate, and 11 undergraduate students.
- Selected 3 runs (High, Middle, Low) and 3 topics (Movie, Shopping, Restaurant) were used.
- The subjects were divided into 3 groups: S_a, S_b, and S_c.



Batch system evaluation measures with NTCIR-5 WEB (269 topics)

	High	Middle	Low
Movie (1196)	S_a	S_c	S_b
Shopping (1296)	S_b	S_a	S_c
Restaurant (1367)	S_c	S_b	S_a



Experimental environment

Document set (NTCIR-5 WEB) approximately 100 million webpages (1.36TB)

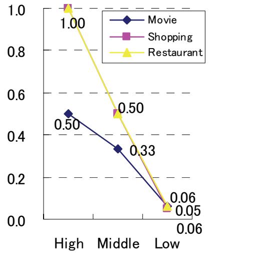
Topics (NTCIR-5 WEB) 269 topics

Run results (NTCIR-5 WEB Rawdata) 63 runs

Searcher (subject)

Example of a selected topic (Shopping)

```
<TOPIC>
<NUM>1296</NUM>
<TITLE>Seiyu, online supermarket</TITLE>
<DESC>I want to visit to Seiyu's online supermarket page.</DESC>
<BACK>I would like to go to shopping at Seiyu's online supermarket.</BACK>
<RELE>Seiyu's online supermarket page in the official Seiyu site is relevant.</RELE>
</TOPIC>
```



RR evaluation measures with selected 3 topics

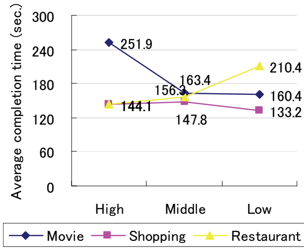
Experiment Procedure

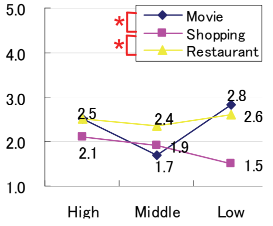
- For each topic, the subjects were instructed to explore the Web, which was in reality NW1000G-04 dataset, through our Web-based interface that presents them a ranked documents list, which corresponds to the assigned system.
- Web interface are similar to usual search engine result pages.
- The subjects were instructed to add a relevant page to bookmark, if they found one. Then, the task for that topic ended.
- At the end of each topic and all the topics, we asked several questions on their perceptions on the systems and topics.

Experiment Results

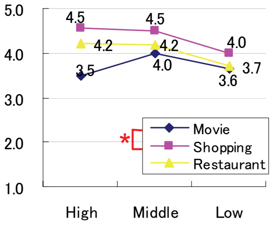
- Completion time: The average search completion time among systems and topics was non-significant.
- Subjective evaluation: The subjects were asked several 5-point scale questions about the systems after the completion; (a) Difficulty, (b) Satisfaction, (c) Confidence, and (d) Appropriateness.

  → No significant difference was observed among system comparisons. A few subjective evaluation among topics are significantly different.
- Comparison with official assessment: At relaxed level, the relevant pages found by the subjects agree with the official assessments in NTCIR-5 WEB in 90.3% in Movie topic, 80.6% in Shopping topic, and 83.9% in Restaurant topic.

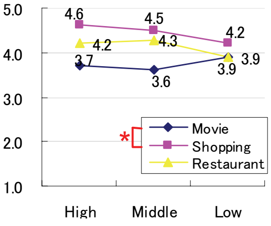




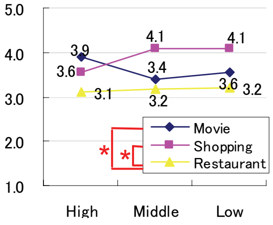
Difficulty



Satisfaction



Confidence



Appropriateness

Conclusion

- Our experiment showed that, in the case of NTCIR-5 WEB Navigational Retrieval task, batch system performance measures (nDCG/MRR/Prec@10) did not match with users' performance evaluations.
- Our simple approach can be easily taken without effective retrieval systems and additional relevance judgements, though the approach itself has a limitation on an interactive data from user experiments.
- In the future work, we will confirm and report on the users' Web navigation characteristics in the experiment.

References

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

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

[3] K. Oyama, et al. Overview of the NTCIR-5 WEB navigational retrieval subtask 2 (Navi-2). In *Proceedings of NTCIR-5 Workshop Meeting*, pp.423–442, 2005.

[4] M. Takaku, et al. Building a Terabyte-scale Web Data Collection "NW1000G-04" in the NTCIR-5 WEB Task. NII Technical Report, NII-2006-012E, 8p, 2006. http://research.nii.ac.jp/TechReports/06-012E.html

Masao Takaku* (Research Organization of Information and Systems)
Yuka Egusa (National Institute for Educational Policy Research)
Hitomi Saito (Aichi University of Education)
Hitoshi Terai (Nagoya University)

*Contact: masao@nii.ac.jp
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan

EVIA-11