

## Overview of WEB Task at the Fourth NTCIR Workshop

Koji Eguchi<sup>†</sup> Keizo Oyama<sup>†</sup> Akiko Aizawa<sup>†</sup> Haruko Ishikawa<sup>†</sup>

<sup>†</sup> National Institute of Informatics  
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan  
{eguchi, oyama, akiko, haruko}@nii.ac.jp

### Abstract

*This paper gives an overview of the WEB Task at the Fourth NTCIR Workshop ('NTCIR-4 WEB') conducted from 2003 to 2004. Through the NTCIR-4 WEB, we investigated the evaluation methods used to measure some tasks of Web information access, such as information retrieval, information classification, and information extraction. We used a 100-gigabyte document dataset that was mainly gathered from the '.jp' domain. Some evaluation measures were applied to individual system results submitted by the participants, and the evaluated values from various viewpoints were compared.*

**Keywords:** Web Information Retrieval, Evaluation Methods, Test Collections.

### 1 Introduction

The Web provides information in all areas of human endeavor. Web information access systems such as search engines provide the necessary means to access the information on the Web. However, evaluation of the effectiveness of such systems has been far from easy, for largely technical reasons. Evaluation workshops and test collections are the most likely means of resolving such problems and should be suitable for the Web.

TREC Web Tracks<sup>1</sup> are well-known evaluation workshops that have the objective of researching the retrieval of large-scale Web document data. Past TREC Web Tracks have used datasets extracted from 'the Internet Archive'<sup>2</sup> or pages gathered from the '.gov' domain as document sets. They only assessed the relevance only on information given in English. The Web Retrieval Task at the Third NTCIR Workshop ('NTCIR-3 WEB')<sup>3</sup> was another evaluation task that used 100-gigabyte and/or 10-gigabyte document

datasets that were mainly gathered from the '.jp' domain. In the NTCIR-3 WEB, relevance judgment was performed on the retrieved documents written in Japanese or English, partially considering hyperlinks.

To further investigate evaluation models for measuring the effectiveness of Web search engine systems, we conducted the WEB Task at the Fourth NTCIR Workshop ('NTCIR-4 WEB') from 2003 to 2004, focusing on Web information access techniques, such as information retrieval, information classification, and information extraction. This paper gives short descriptions on task design of the NTCIR-4 WEB.

### 2 Overview of Task Descriptions at NTCIR-4 WEB

The WEB Task at the Fourth NTCIR Workshop (NTCIR-4 WEB) attempted to push ahead research of information access systems for large-scale Web documents, making use of the experiences of the NTCIR-3 WEB. The organizers investigated actual use of the Web from various viewpoints, and designed the following subtasks to evaluate the required fundamental techniques.

**Subtask A:** Informational Retrieval Task 2

**Subtask B:** Navigational Retrieval Task 1

**Subtask C:** Geographic Information Task 1<sup>4</sup>

**Subtask D:** Topical Classification Task 1

The names of the Informational Retrieval Task and the Navigational Retrieval Task were derived from Broder's taxonomy [1], as were mentioned in TREC Web Tracks. The Informational Retrieval Task was designed to evaluate the effectiveness of search engines from the viewpoint of topic relevance<sup>5</sup>, considering hyperlink relationship and content duplication.

The Navigational Retrieval Task was designed to evaluate the effectiveness of the search engines, assuming that a user is motivated to find a small num-

<sup>4</sup>This subtask was organized by Masatoshi Arikawa and Takeshi Sagara at the University of Tokyo.

<sup>5</sup>This subtask was based on the 'Survey Retrieval Task' and the 'Target Retrieval Task' at NTCIR-3 WEB.

<sup>1</sup><http://es.csiro.au/TRECWeb/>

<sup>2</sup><http://www.archive.org/>

<sup>3</sup><http://research.nii.ac.jp/ntcweb/>

**Table 1. Fundamental statistics of the document sets**

Statistics of NW100G-01	
(1-1) # of crawled sites *	97,561
(1-2) max. # of pages within a site	1,300
(1-3) # of crawled pages **	11,038,720
(1-4) # of pages for searching	15,364,404
(1-5) # of links connected from (1-3)	78,175,556
(1-6) # of links connected from (1-3) to (1-4) ***	64,365,554

(\*) Aliased sites are not included.  
 (\*\*) *i.e.*, # of pages included in the document data for provision of access. Aliased sites are not included.  
 (\*\*\*) *i.e.*, # of pages included in the document data for reference. The existence of the other pages, *i.e.*, (1-6)-(1-5) or (2-6)-(2-5), could not be confirmed.

ber of typical Web pages of a known item, such as a person, shop, restaurant or facility.

The Geographic Information Task investigated the feasibility of evaluating techniques that extract geographical descriptions from the Web pages relevant to both location and one, or more, descriptive elements.

The Topical Classification Task attempted to evaluate techniques for supporting users' browsing processes by means of classification-based output presentation, such as using clustering techniques, while assuming that users submit very short queries having ambiguity<sup>6</sup>.

### 3 Document Sets

The document sets were explicitly specified for the test collections. In the NTCIR-4 WEB, we used 'NW100G-01' data that were constructed at the Web Retrieval Task at the Third NTCIR Workshop (NTCIR-3 WEB), as the document set. The NW100G-01 is composed of the document data gathered from the '.jp' domain. We also provided a separate list of documents that were connected from the individual documents included in the NW100G-01 data, but not limited to the '.jp' domain. These two datasets were used for processing at the NTCIR-4 WEB. The Geographic Information Task and the Topical Classification Task used 'target datasets' that were separately defined as subsets of the NW100G-01 data. Those subsets, while comparatively small-scaled, were suitable for the designed task.

Fundamental statistics of the document sets are shown in **Table 1**. The crawling strategy is described in Reference [2]. We stored the NW100G-01 data in a hard disk drive and delivered it to each participating group. For the purpose of handling the NW100G-

<sup>6</sup>This subtask was based on the 'Search Results Classification Task' at the NTCIR-3 WEB, which was proposed as a pilot study and discussed as one of the 'Optional Tasks' at the NTCIR-3 WEB; however, no classification results were submitted on time.

01 data, the computer resources at the 'Open Laboratory' located at the National Institute of Informatics, Japan were available only for those participants who requested to use them<sup>7</sup>.

### 4 Conclusion

We briefly described the task design of the NTCIR-4 WEB. The details of each subtask are described in References [3, 4, 5, 6] included in this volume.

Ohtsuka *et al.* proposed a user-oriented criterion for evaluating Web search systems, considering users' search behavior. As organizers of the NTCIR-4 WEB, they attempted to evaluate the Informational Retrieval Task using a part of the data of submitted run results and the topics [7].

The Web test collections that we have developed through the NTCIR-4 WEB will be available for re-search purposes<sup>8</sup>.

### References

- [1] A. Broder. A taxonomy of web search. *SIGIR Forum*, 36(2):3–10, 2002.
- [2] K. Eguchi, K. Oyama, E. Ishida, N. Kando, and K. Kuriyama. Evaluation methods for web retrieval tasks considering hyperlink structure. *IEICE Transactions on Information and Systems*, E86-D(9):1804–1813, 2003.
- [3] K. Eguchi, K. Oyama, A. Aizawa, and H. Ishikawa. Overview of the Informational Retrieval Task at NTCIR-4 WEB. In *Proceedings of the Fourth NTCIR Workshop on Research in Information Access Technologies: Information Retrieval, Question Answering and Summarization*, Tokyo, Japan, 2004.
- [4] K. Oyama, K. Eguchi, H. Ishikawa, and A. Aizawa. Overview of the NTCIR-4 WEB Navigational Retrieval Task 1. In *Proceedings of the Fourth NTCIR Workshop on Research in Information Access Technologies: Information Retrieval, Question Answering and Summarization*, Tokyo, Japan, 2004.
- [5] M. Arikawa, T. Sagara, K. Noaki, and H. Fujita. Preliminary workshop on evaluation of geographic information retrieval systems for web documents. In *Proceedings of the Fourth NTCIR Workshop on Research in Information Access Technologies: Information Retrieval, Question Answering and Summarization*, Tokyo, Japan, 2004.
- [6] K. Eguchi. Overview of the Topical Classification Task at NTCIR-4 WEB. In *Proceedings of the Fourth NTCIR Workshop on Research in Information Access Technologies: Information Retrieval, Question Answering and Summarization*, Tokyo, Japan, 2004.
- [7] T. Ohtsuka, K. Eguchi, and H. Yamana. An evaluation method of web search engines based on users' sense. In *Proceedings of the Fourth NTCIR Workshop on Research in Information Access Technologies: Information Retrieval, Question Answering and Summarization*, Tokyo, Japan, 2004.

<sup>7</sup>At the NTCIR-3 WEB, all participants were only allowed to process the NW100G-01 data within the 'Open Laboratory' located at the National Institute of Informatics. However, we changed the method for providing the document sets at the NTCIR-4 WEB as mentioned above.

<sup>8</sup><http://research.nii.ac.jp/ntcweb/>