Introduction

Formal Run Results of Navigational Retrieval Subtask in NTCIR-4 WEB

NTCIR-4 WEB Task Organizers
ntcadm-web@nii.ac.jp

Jun 2, 2004

Abstract

This article gives additional information on formal run results of the Navigational Retrieval Subtask 1 in the WEB Task at the Fourth NTCIR Workshop (NTCIR-4 WEB). The run results and this article are available on the CD-ROM and in the NTCIR Web site[*].

Please pay attention to the following points in reading 'Evaluation Result Sheets' and 'Summary of Evaluation Results'.

Evaluation Result Sheets

We evaluated each run result on two different document sets and two different relevance levels, resulting four combinations.

Document sets

(RL-1) Document set with and without delivered page data
The document set which is originally defined by the task description. It includes all the documents whose identifiers are included in 'targetlist' file, i.e., not only documents whose page data are delivered to the participants, but also documents which have only in-links from one or more of them and are fetched and stored by the organizers, are the targets of retrieval.

(RL-2) Document set with delivered page data
An additional document set for comparison. It includes all the documents whose identifiers are included in 'doclist' file, i.e., only documents whose page data are delivered to the participants. It is a subset of (1).

Relevance levels

(RL-1) Rigid (left half of each sheet)
Documents with assessment A ('relevant') are regarded as relevant documents.

(RL-2) Relaxed (right half of each sheet)
Documents with assessment A ('relevant') and B ('partially relevant') are regarded as relevant documents.

---

[*] http://research.nii.ac.jp/ntcir/

We delivered 300 topics to the participants and assessed 144 before the evaluation. For each document set, we selected such topics that at least one relevant document at the 'rigid' relevance level was included in each of their pools. Consequently, we used 87 topics (TS-1) for (DS-1) and 72 topics (RS-2) for (DS-2).

As the evaluation measures, we calculated DCG and MRR at top ranked 10 documents for each of above mentioned four combinations, averaging over (TS-1) and (TS-2) respectively. Gains used in the DCG calculation are: $(G_A, G_B) = (3, 0)$ for (RL-1), and $(G_A, G_B) = (3, 2)$ for (RL-2).

Although many topics have multiple relevant documents for each, most of them are either duplicated web pages or closely linked web pages. Therefore, for each run, the first retrieved relevant document has importance and the others have little. Because duplication and link relation are not considered in this evaluation, appropriateness of DCG values as the system effectiveness is left to be investigated. However, because only the first retrieved relevant document is used in MRR, the appropriateness is the same regardless of considering duplication or link relation.

We also included two graphs for each of above mentioned four combinations:

- Difference from the top 20th run in reciprocal rank of top ranked relevant document per topic
  For each topic, difference of the run from the top 20th run in the reciprocal ranks of the relevant documents retrieved first respectively is plotted. Note that the top 20th run is determined per topic.
- Cumulative number of topics for which one or more relevant documents were retrieved
  The number of topics, each of which one or more relevant documents were retrieved for above a given rank, is plotted.

Summary of Evaluation Results

Runs by the participants were selected, one per participant, according to the priority specified in the system description form. Runs by the organizers (run-IDs' prefixed with 'ORGREF-') were selected, two for baselines as content only IR systems (*-NMZ-AND and *-OT-DT), three for different types of link and anchor usage (*-OT-D-LF2, *-OT-DT-LB2, and *-AT40-P1).

Above mentioned graphs 'Cumulative number of topics for which one or more relevant documents were retrieved' of the selected runs are plotted together.

Summary of Topics

Run results per topic are summarized in two types of graphs as follows.

- Number of runs that retrieved relevant documents per topic
  The cumulative number of runs by which one or more relevant documents were retrieved, each at rank 5, 10, 20, and 100, is plotted per topic.
- Average reciprocal rank over runs per topic
  The average over all runs per topic in reciprocal rank of relevant documents, each of which was retrieved first by each run, is plotted.