Introduction

Formal Run Results of Informational 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 Informational Retrieval Subtask 2 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[1]. Please pay attention to the following points in reading `Evaluation Result Sheets.'

Evaluation Result Sheets

We evaluated each run result under two user behavior models and two different relevance levels, resulting four combinations. All the run results were evaluated using the document set that was originally defined by the task description. This includes all the documents whose identifiers are included in `targetlist' file, i.e., not only documents whose page data were delivered to the participants, but also documents that had only in-links from one or more of them and were fetched and stored by the organizers.

User behavior models

(UM-1) *`Survey'-type* (the first page)

Is assuming the model where the user attempted to comprehensively find documents relevant to his/her information needs.

(UM-2) *`Target'-type* (the second page)

Is assuming the model where the user requires just one or only a few relevant documents at the highly ranked documents.

Relevance levels

(RL-1) *Rigid* (left half of each sheet)

Documents with assessment S (`highly relevant') or A (`fairly relevant') are regarded as relevant documents.

(RL-2) *Relaxed* (right half of each sheet)

Documents with assessment S (`highly relevant'), A (`fairly relevant') or B (`partially relevant') are

[1] http://research.nii.ac.jp/ntcweb/

regarded as relevant

We delivered 153 topics to the participating groups; however then we discarded inappropriate topics and we obtained resulting 128 topics. Consequently, we used 35 topics using document pools from highly ranked 100 documents —assuming UM-1 —, and 75 topics using document pools from top 20 documents —assuming UM-2—. These 35 topics were also used for the evaluation under UM-2; therefore 80 topics could be used for the evaluation under UM-2.

As the evaluation measures, we calculated evaluation values based on recall and/or precision,[2] DCG and MRR for each of above mentioned four combinations. Gains used in the DCG calculation are: (G_S, G_A, G_B) = (3, 2, 0) for (RL-1), and (G_S, G_A, G_B) = (3, 2, 1) 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. The duplication and link relation are not considered in this evaluation; however, these will be discussed continuously as a future issue.

Comparable Graphs and Charts of Evaluation Results

Comparable graphs, such as recall-precision curves and DCG curves, and charts of selected runs are shown in the overview paper included in the working-note proceedings.

Remarks

Details of the evaluation methods are described in the overview paper. Additional evaluation results will be put on the NTCIR Web site[3].

[2] We used `trec_eval', a program that evaluates TREC results. This is available at <ftp://ftp.cs.cornell.edu/pub/smart/trec_eval.v3beta.shar>.

[3] http://research.nii.ac.jp/ntcweb/