TREC12 Web and Interactive Tracks at CSIRO

Nick Craswell[1], David Hawking[1] and Trystan Upstill[2]
nick.craswell@csiro.au, david.hawking@csiro.au and trystan.upstill@cs.anu.edu.au

Alistair McLean[1], Ross Wilkinson[1] and Mingfang Wu[1]
alistair.mclean@csiro.au, ross.wilkinson@csiro.au, mingfang.wu@csiro.au

[1]CSIRO ICT Centre, Australia

[2]Department of Computer Science, CSIT Building, ANU

Canberra, ACT 0200, Australia

1. Introduction

This year, CSIRO teams participated in all three tasks of the web track; these being: the automatic topic distillation task, the home/named page finding task and the interactive topic distillation task. This paper describes our approaches, experiments and results. The following section describes our experiments in the two automatic tasks, and Section 3 describes our experiment in the interactive task.

2. The web track

CSIRO submitted a total of 10 runs to the non-interactive portion of the 2003 Web Track - 5 runs for home/named page finding and 5 runs for Topic Distillation. The runs are labeled csiro03[TYPE][RUNID], where TYPE is ``ki" for known item runs and ``td" for topic distillation runs.

This year we focused on tuning Okapi BM25 and Web evidence parameters. Our home/named page finding submissions use tunings computed for both home and named page finding, and evaluate two run combination methods. Our topic distillation submissions are tuned for home page finding only and test whether the Web evidence evaluated is useful, and whether the use of stemming improves performance.

We did not incorporate PageRank or simple indegree this year because of previously observed poor performance for named page finding and homepage finding. Instead our query-independent Web evidence included URL length and two important sub-types of indegree (off-site and on-site).

Throughout our experiments we tuned Okapi BM25 (through the k_1 and b parameters), anchor-text weighting and other query independent Web evidence. The parameters were tuned using a hill climbing algorithm, with complete exploration of 2 parameters at a time, on a grid computer consisting of cluster of 20 dual processor Intel Xeon machines.

2.1. Home/named page finding

We submitted runs based on three tunings (for a home page task, a named page task and both at the same time), and evaluated two combination methods. We trained using last year's .GOV named page finding query/result set, and using a home page finding training set derived from a first .GOV resource listing.

We submitted runs tuned for both home page and named page finding at the same time (csiro03ki01), tuned for named page finding only (csiro03ki02) and tuned for home page finding only (csiro03ki03). We also submitted two combinations of these runs. The first was an interleaved run (csiro03ki04 -- interleaving csiro03ki02 and csiro03ki03), and the second a run that summed scores achieved in both rankings (csiro03ki05). A summary of our home/named page finding submissions, and their retrieval effectiveness is presented in Table 1.