Plans for the TREC-9 Web Track

David Hawking
CSIRO Mathematical and Information Sciences,
Canberra, Australia
David.Hawking@cmis.csiro.au

In recent years, TREC has broadened its scope to include many more facets of the Web searching process. In TREC-8 (1999), the Web special interest track evaluated link-based retrieval methods investigated differences between Web and traditional TREC ad hoc documents, and studied efficiency and effectiveness tradeoffs on large data sets. In addition, although neither used Web data, both the Cross-Lingual track and the Question & Answer track studied issues of considerable importance to everyday Web search.

In TREC-9, the main Web track task will use a larger set of Web documents than last year and will use search topics derived from search engine logs. This task will be the closest approximation in TREC-9 to the traditional TREC Ad Hoc retrieval task.

Main Web Task

This Web task will use a 10 gigabyte collection of web pages (to be known as WT10g) and a set of topic descriptions which will be reverse-engineered from queries taken from Web search engine logs. Submissions will be evaluated by NIST assessors in the time-honoured way but it is possible that the assessors will also try to identify the "best" answer as well as all relevant answers. Some evaluation measures may also take into account the indirect value of pages which, while not themselves relevant, may link to relevant pages.

Peter Bailey is coordinating the definition and construction of the WT10g with input from colleagues at ACSys and UVA. The task is to select documents for this collection in such a way as to satisfy as many as possible of the following desiderata.

The collection should:

1. contain content likely to be relevant to a cross-section of Web search topics;
2. contain an interesting, or at least realistic, population of inter-site links;
3. contain few binary or non-English documents (to remove unnecessary barriers to participation). However, site home pages which happen to be empty or highly graphical will not be removed even if they meet the rejection criteria.;
4. represent the union of a realistic set of substantially complete server holdings (to enable realistic experiments by groups interested in distributed information retrieval). We will endeavour to make the distribution of server sizes the same as that of the VLC2 collection.

Other issues to be resolved are how to deal with duplicate pages and with generated content.

NIST assessors will develop topics by taking queries from public Web search engine logs and running pilot searches over the test corpus. It is envisaged that the actual Web query will be used as the Title field and that Description and Narrative fields will be reverse engineered by the assessors.

Note that the main Web task will be a straightforward ad hoc retrieval task but that participants will be able to conduct their own distributed IR and link-based method experiments by submitting multiple runs. It is expected that run submission questionnaires will be modified to identify what experiments were being conducted.

I understand that the TREC-9 Interactive Track may use the WT10g in a Q&A context.

Connectivity Data / Ability to Follow Links

Nick Craswell has computed a complete connectivity matrix for the VLC2 collection. We will make a WT10g subset (with appropriate DOCNO translation) and distribute it either by FTP or on CD-ROM.

Nick has also implemented a means by which document links can be followed within the test collection. It works by translating URLs to DOCNOs and using the browser proxy mechanism. Hopefully, this prototype can serve as the basis for a robust and portable version suitable for use if required in the Interactive Track.

Large Web Task

It is likely that there will again be a Large Web retrieval task similar to last year's. At this stage it is uncertain whether the data set will be the 100 gigabyte VLC2 or whether a 300 gigabyte superset will be constructed. It is also unclear whether relevance assessments will be made by participants or whether participants will contribute to the cost of centrally performed assessments. Resolution of these issues is unlikely until after the WT10g collection is defined and distributed.

Participation

To participate in the Web Track, unless you have already done so, you should:

1. Register for TREC-9 and put your name down for the Web Track;
2. Mail me (`david.hawking@cmis.csiro.au`) and ask to be added to the webtrax mailing list.

Access to Data Sets

The Web test collections (VLC2, WT2g, WT10g) are distributed by ACSys/CSIRO. To obtain these collections your organisation must sign an agreement with us and pay a contribution to the cost of preparing and distributing the data. The agreement is similar to the TREC data permission forms but also requires you to delete documents if requested to do so.

You can obtain copies of the agreement and other information and order data sets from the website: `http://pastime.anu.edu.au/WAR/WT/`. To access this site, you can use either the TREC-9 username/passwordcombination or a web-track specific combination.