Task Descriptions: Web Track 2003

Nick Craswell and David Hawking
CSIRO ICT Centre
Canberra, Australia
nick.craswell@csiro.au and david.hawking@csiro.au

Ross Wilkinson and Mingfang Wu
CSIRO ICT Centre
Melbourne, Australia
ross.wilkinson@csiro.au and mingfang.wu@csiro.au

October 30, 2003

Part I
Non-interactive Experiments

1. Introduction

This Year's Aims

- To investigate methods for effective topic distillation: Finding a set of the best home pages, given a broad query.
- To investigate methods for effective navigational search, with a mixture of home page and named page queries: Finding a particular page desired by the user.
- To increase the available queries/judgments for the .GOV test collection.

Participants are welcome to explore other Web retrieval issues such as distributed IR, queries with misspellings, efficient indexing etc within the context of these experiments.

2. Dataset

The corpus for both tasks is the .GOV test collection, distributed by CSIRO. Documents include the information returned by the http daemon (enclosed in DOCHDR tags) as well as the page content.

The crawl is recent (start of 2002). It is the sort of crawl which might be used by a real .gov search service: breadth first, stopped after the first million

html pages and including (the extracted plain text of) an additional 250,000 non-html pages (doc, pdf and ps). Text is provided for convenience. NIST assessors will view original (binary doc,pdf,ps,gif,jpg) files when judging.

3. Topic distillation task

Topic distillation involves finding a list of key resources for a particular topic. In this year's task we are concentrating solely on websites as resources. The task is to find as many different websites (represented by their entry pages) as possible within the first ten results.

For the topic, 'science', the following websites might be considered key resources:

```
www.nsf.gov/                        National Science Foundation
science.nasa.gov/                   Science @ NASA
www.science.gov/                    Government Science Portal
www.house.gov/science/welcome.htm House Committee on Science
```

To be judged a "key resource", the page returned should be a good entry point to a website which:

- Is principally devoted to the topic,
- Provides credible information on the topic, and
- Is not part of a larger site also principally devoted to the topic

For the 'science' topic, the page 'www.house.gov' fails the first test while the page 'www.nsf.gov/home/bio/' fails the third. Hopefully within the .gov domain, it will be hard to find sites which fail the second test! NIST will develop topics for .GOV. Example topic format:

```
<top>

<num> Number:
<title> science

<desc> Description:
Find key government websites (represented by their home page)
on the subject of 'science'.

</top>
```

The title field only should be supplied to your system as the query.

Systems will be judged according to how many good answers they find in the top ten results (the first page returned by a typical Web search system). Likely measures are precision at 10 and average precision at 10.

4. The home/named page finding task

Users sometimes search for a page by name. In such cases, an effective search system will return that page at or near rank one.

This year's task involves a mixture of tasks from two previous years: home page finding and named page finding. In both cases, there is only one target page and user queries are often the name of the page. The difference is that home page finding queries are restricted to home pages: 'Internal Revenue Service' → 'www.irs.gov', while named page finding may involve pages which are not home pages 'passport application form' → 'travel.state.gov/dsp11.pdf'. Some search/ranking metrics will be useful for both types of query, while others will only be useful for one.

NIST will devise the mixed set of queries for .GOV. A minimal amount of judging will be required to determine if the URLs of documents returned by participants are in fact equivalent to the answer originally chosen. For example, if the page is available at 2 different URLs, both would be considered correct answers.

Systems will be compared on the basis of the rank of the first correct answer. Likely measures include mean reciprocal rank of first correct answer and success rate at N (percentage of cases in which the correct answer or equivalent URL occurred in the first N documents).

No manual or interactive query modification is permitted in this task.

5. Indexing Restrictions

There are none. You may index all of each document or exclude certain fields as you wish.

6. Submissions and Judgments

All submissions are due at NIST on or before 6 August 2003.

Submission information:

Topic distillation: Submit up to 5 runs. For each query, list up to 1000 (the top 1000) results. Check your results using `check_web.pl` (available from `http://trec.nist.gov/`)

Home/named page finding: Submit up to 5 runs. For each query, list up to 50 (the top 50) results. Check your results using `check_web.pl` (available from `http://trec.nist.gov/`)

The result format is:

```
topic-id Q0 docno rank sim tag

  topic is the topic number,
  Q0    is the literal 'Q0',
  docno is the document id taken from the DOCNO field of the text,
```

```
rank  is the rank assigned to the document,
sim   is the similarity computed between the document and the topic,
tag   is the run tag.
```

It is likely that NIST will accept up to 5 official submissions for each task, but the number of fully judged runs per group will depend upon the number of submissions, the degree of overlap and the judging resources available. Hopefully it will be possible to judge two topic distillation runs and two home/named page runs per group.

All judging will be performed by NIST assessors.

Judgments will be binary. Key resource OR Not key resource. Home/named page OR Not home/named page.

Judgments will be made on the basis of the text within the document, its URL and in the case of topic distillation the pages it links to (particularly those on the same site).

Part II
Interactive Experiments

1. Motivating principles

This year the interactive track will become a sub-track of the web track. One of the web track tasks, the topic distillation, has been selected as the interactive track task. However, the interactive sub-track will focus on the human participation in topic distillation. Virtually any kind of user studies on topic distillation would be acceptable. A pre-defined protocol will be provided for the balanced studies that the past interactive tracks have used.

2. Tasks

An information searcher may often need to construct a list of resources on a topic for him/her own learning or for other people. Such a resource list could be manually constructed, such as that in Yahoo!, or automatically constructed for a user-defined topic by a topic distillation algorithm.

In the interactive sub-track, a searcher will be asked to construct such a resource list on a broad topic through interaction with an information access system. A typical task statement (or instruction) given to the searcher would look like this:

> *"Your task is to construct a learning resources for a class of 16-year-old secondary school pupils on [topic x]. Your learning resource should include the good main pages that point to websites that together cover all the major aspects of [this topic]."*

The following eight search topics are selected from the topic set as used by the web topic distillation task, and are rephrased according to the format of the above task statement.

1. **Title:** cotton industry
Search task:
You are to construct a resource list for high school students who are interested in cotton industry, from growing, harvesting cotton and turning it into cloth.

2.**Title:** folk art folk music
Search task:
Assume that you are an art teacher of a high school. You are about to introduce your students to U.S. folk art and folk music. Please prepare a list of bookmarks for your students for study materials.

3. **Title:** children's literature
Search task:
The teachers from your local primary school are spending a lot of their time on the web to search for materials on children's literature. Please help the teachers by setting up a children's literature web guide which points to useful websites for young readers/writers.

4. **Title:** wireless communications
Search task:
You are invited to give a presentation on wireless communication to university students. Please prepare a list of bookmarks as a handout to your audience. The bookmarks should cover information on existing and planned uses, research/technology, regulations and legislative interest.

5. **Title:** arctic exploration
Search task:
Assume that you are a high school student and working on an arctic exploration project. You are asked to collect some resources from the web for your project team on what kinds of exploration of the arctic are underway, especially of glaciers and ice.

6. **Title:** weather hazards and extremes
Search task:
Assume that you are a high school student and working on a project regarding the study of natural/weather hazards and extremes. You are asked to collect some resources from the web for your project team.

7. **Title:** electric automobiles

Search task:
You are going to give a seminar on the progress in producing/developing electric automobiles, and you will mention some online resources on this topic. Please prepare a list of bookmarks as a handout to your audience.

8. **Title:** Bilingual education
Search task:
You are a volunteer of your local community. You are asked to help to create a guide to all online information on bilingual education that may be of interest to your local residents.

3. Data collection and search systems

The interactive track will use the same .GOV web collection created for the TREC 2003 web track.

Panoptic at NIST

NIST will provide access to its server with the "Panoptic search engine" and the "Panoptic topic distillation engine" (both of them are accessible through `http://ir.nist.gov/`). You can also find the link to the help page there. A short description of each search algorithm is also available.

In order to be consistent with the web topic distillation task, all searches and browses are restricted within the .GOV collection. Thus, the Panoptic topic distillation engine has been configured so that all hits, and all links in hit pages, point back into the .GOV collection. (Please note this has some consequences that some links don't work, some images are missing, and this affects some pages more than others.)

Advanced usage

- You can request a page by its URL,

    `http://ir.nist.gov/search/gov.cgi?url=http://trec.nist.gov/`

- or by its docid:

    `http://ir.nist.gov/search/gov.cgi?id=G01-01-0000000`

XML search interface

If you have your own interface and just want to feed queries to Panoptic and get XML output, you can search using the search-xml.cgi script: `http://ir.nist.gov/search/search-xml.cgi?query=wireless+communication&collection=gov` The CGI query syntax is the same as for the standard search interface. Panoptic has a number of fancy query operators, and if you want to know how

to feed them to the XML script, give them to the standards interface and look at the URL of the results page.

Participants are free to use any appropriate search engine, but need to make sure that all searches and browses are within the .GOV collection.

4. Experimental protocol

Participants are free to define an experimental protocol that suits their own experiment purpose. For those who want to compare two systems or system variants, they could adopt the following experimental protocol.

Experimental procedure

1. Entry questionnaire
2. Tutorial session
3. Before-search questionnaire per topic
4. Topic search
5. After-search questionnaire per topic
6. After-system questionnaire per system after 4 topics on a system
7. Exit questionnaire

Experimental design

The design is within subjects, and requires 16 subjects for completion. It depends upon dividing the eight topics into two blocks, varying the order of topics within each block.

If we define:

```
B1 = block 1;
B2 = block 2;

a = order 1,
b = order 2,
c = order 3,
d = order 4
```

	a	**b**	**c**	**d**
B1	1234	4321	3142	2413
B2	5678	8765	7586	6857

`Then the first four subjects search as follows:`

S1	System I: B1a	System II: B2a
S2	System I: B2a	System II: B1a
S3	System II: B1a	System I: B2a
S4	System II: B2a	System I: B1a

This pattern is then repeated for each of the remaining 3 topic orders (b,c,d).

Instructions to be given to searchers(during the tutorial session)

In this experiment, your task is to construct a list of key resources on a given topic. This resource list is intended to guide someone who shows interest in that topic to find more information. The goal of this experiment is to determine how well an information retrieval system can help you accomplish your task.

Each of these key resource pages should be a main page of a website which:

1. Is principally devoted to the topic,
2. Provides credible information on the topic, and
3. Is not part of a larger site also principally devoted to the same topic

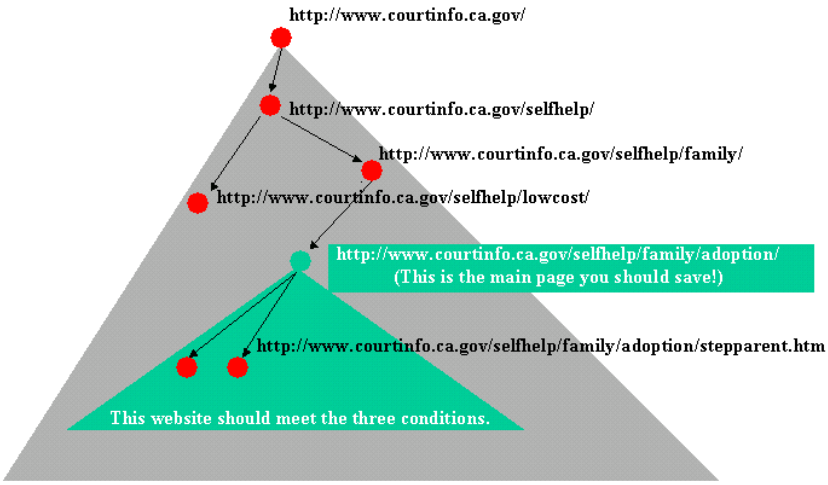
Here is an example:

> **Topic:** Adoption procedures
>
> **Search task:** To construct a resource list for those people who want to adopt a child. Please try to find and save those main pages pointing to websites that together should cover all major aspects on adoption.
>
> For example: the following pages would be regarded as good main pages:
>
> G00-03-2173112 (`http://www.acf.dhhs.gov`) - Administration for Children and Families (USHHS)
> G13-55-1080004 (`http://www.acf.dhhs.gov/programs/cb/dis/afcars/`) - Adoption and Foster Care Analysis and Reporting System (USHHS)
> G07-03-3445073 (`http://www.courtinfo.ca.gov/selfhelp/family/adoption/`)- California Courts Self-Help Center
> G00-98-2804800 (`http://www.mcdss.co.gov`) - Mesa County Dept. of Human Services
>
> The above pages are acceptable because they are the main pages of those websites that meet all three conditions. To decide whether a current page is the main page you should save, you can try to go up or down the links in the current page, and see whether all three conditions are still meet. You may find the up link of a main page may not be principally devoted to the topic, while the down link of

a main page may cover only a part of the topic. For example, the following pages would NOT be regarded as a good main page:

G00-08-0239407 (`http://www.courtinfo.ca.gov/`) - California Courts - The Judicial Branch of California - fails the first condition

G33-75-3683089 (`http://www.courtinfo.ca.gov/selfhelp/family/adoption/stepparent.htm`) - Stepparent Adoptions in California - fails the third condition



Here is a hand-on practice topic:

Topic: intellectual property

Search task: You are a high school social studies teacher, and are about to introduce a module on intellectual property. Prepare a list of resources for the class that define intellectual property, and explain how creators of intellectual property are protected under US laws.

(Some example good pages:

G01-23-1524097 (`http://www.loc.gov/copyright/about.html`) - Welcome to the US Copyright Office
G00-03-2959565 (`http://patents.uspto.gov`) - US Patent and Trademark Office
G01-17-4061425 <DOCHDR> - www.cybercrime.gov/ip.html
G02-48-1320654 (`http://patents.gsfc.nasa.gov`) - NASA Office of Patent Counsel

)

5. Evaluation

For each topic, the following two qualitative measures will be applied and gathered.

> **Accuracy:** The extent of relevance of each page in a resource list to the corresponding search topic. The judgement will be on a seven-point Likert scale.
>
> **Coverage:** The relative coverage of relevant aspects in a resource list. (After all submitted pages per topic are aggregated and judged for the accuracy, then assessors will judge each list for the coverage.) The judgement will also be on a seven-point Likert scale.

The subjective evaluation (such as searchers' satisfaction) will be collected through the After-search questionnaire, After-system questionnaire and Exit questionnaire.

Data to be submitted to NIST

Each site should submit one ascii file only. Each line in the file represents a search topic worked on by a subject even if no DOCNO is saved. If you have 16 subjects and 8 topics, then your file should have 16X8 = 128 lines. Each line should contain the following items with intervening spaces and semicolons as indicated. Since semicolons will be used to parse the lines, they can only occur as specified in the following format:

```
SiteID; SystemID; SearcherID; TopicNum; DOCNOLIST

where:

     SiteID - unique across sites

     SystemID - unique within site to each of your IR systems

     SearcherID - unique within site to each of your subjects

     TopicNum - the topic number as in the guidelines

     DOCNOLIST - a list of TREC DOCNOs as found in the
                    documents, separated by commas.

Sites determine SiteID, SystemID, and SearcherID. They are not
allowed to contain spaces.
```

Sites are not required to submit those data from questionnaires.

6. Schedule

- 30 June 2003 - all resources (search topics and search engines) will be available, and Guidelines will also be completed.
- 31 August 2003 - All runs should be sent to Ian Soboroff for the judgment of accuracy and coverage.
- 8 September 2003 - Each participating group will submit a paragraph to Ross Wilkinson including a brief description of their studies.