Overview of the TREC 2003 Web Track

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

March 22, 2004

Abstract

The TREC 2003 web track consisted of both a non-interactive stream and an interactive stream. Both streams worked with the .GOV test collection. The non-interactive stream continued an investigation into the importance of homepages in Web ranking, via both a Topic Distillation task and a Navigational task. In the topic distillation task, systems were expected to return a list of the homepages of sites relevant to each of a series of broad queries. This differs from previous homepage experiments in that queries may have multiple correct answers. The navigational task required systems to return a particular desired web page as early as possible in the ranking in response to queries. In half of the queries, the target answer was the homepage of a site and the query was derived from the name of the site (Homepage finding) while in the other half, the target answers were not homepages and the queries were derived from the name of the page (Named page finding). The two types of query were arbitrarily mixed and not identified.

The interactive stream focused on human participation in a topic distillation task over the .GOV collection. Studies conducted by the two participating groups compared a search engine using automatic topic distillation features with the same engine with those features disabled in order to determine whether the automatic topic distillation features assisted the users in the performance of their tasks and whether humans could achieve better results than the automatic system.

Part I
Non-interactive Experiments

1 Introduction

The non-interactive stream of the TREC 2003 Web Track centred on two tasks: a topic distillation task and a navigational task. The tasks use the 18 gigabyte, 1.25 million document partial crawl of the .gov domain, distributed on CD-ROM as the .GOV collection[1]. A full description of this year's track guidelines is available in a separate document in these proceedings.

[1] Seehttp://es.csiro.au/TRECWeb/

2 Tasks

This year's tasks represent two types of search where it is important for the system to be able to return homepages.

A homepage is designed to be the main page of a site. The homepage is important because it is often the first page users will see. It provides an introduction to the site, who created it and what it contains. It usually links to other pages in the site and provides access to other site functions such as search. The homepage URL is often given as the URL of the whole site, as in 'the TREC site is at `http://trec.nist.gov/`'.

In previous years Track participants have developed effective methods for homepage finding. Link-based ranking methods including anchortext propagation are useful, because homepages tend to have higher inlink counts than non-homepages. URL-based ranking is also useful, since homepages tend to have short URLs. This year, such evidence might be expected to be useful in processing topic distillation queries and for some of the navigational queries. In this year's navigational task, the hompepage queries were mixed with an equal number of queries designed to find named pages which were not site homepages.

2.1 Topic Distillation Task

The topic distillation task involves finding relevant homepages, given a broad query. The need underlying the query 'cotton industry' might be 'give me an overview of `.gov` sites about the cotton industry, by listing their homepages'. See Figure 1.

This differs from an adhoc-style interpretation 'give me all pages in `.gov` about the cotton industry'. Adhoc querying would facilitate direct access to a much larger number of pages, but topic distillation gives a better overview of which sites exist and therefore which government labs, groups and programs have sites.

It also differs from past Web track homepage finding tasks in that queries do not identify a specific site. To illustrate the difference, a homepage finding query in the 'cotton industry' area might be 'cotton pathology research unit'.

The topics were numbered 1–50. A good homepage will correspond to a site 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

This requires judges to understand the structure of the site in question and the quality of information offered, and identify its homepage. We have more emphasis on homepageness than in last year's topic distillation task, and also have used broader queries to ensure that at least some sites exist.

Because many topics had less than 10 results, we abandoned the precision at 10 measure (P@10). The main measure was R-Precision (P@n where n is the number of relevant documents for the current topic).

2.2 Navigational Task

The navigational task is also known as the 'home/named page finding task'. Each query involves finding a particular page, which is a homepage in 50% of queries (participants did not know which queries were for homepages and which were for non-homepages). The query asks for the page by name. For example, when looking for the homepage `http://www.tva.gov` the user might type the query 'TVA-Tennessee Valley Authority'. When looking for the page `http://www.usdoj.gov/crt/ada/enforce.htm`, they might type the query 'ADA Enforcement'. See Figure 2.

Within the framework of this task, a number of research questions can be addressed, including:

1. Do systems tuned for homepage finding also work well on the named page finding task?

```
<top>
<num> Number: TD7
<title>cotton industry</title>
<desc>Description:
Where can I find information about growing, harvesting cotton
and turning it into cloth?
</top>

-------
Answers
-------

Cotton Pathology Research Unit
 cpru.usda.gov/

FAS Cotton Group
 ffas.usda.gov/cots/cotton.html

The Western Cotton Research Laboratory
 nps.ars.usda.gov/locations/locations.htm?modecode=53-44-05-00

Office of Textiles and Apparel
 otexa.ita.doc.gov/

U.S. Cotton Data Sets
 wizard.arsusda.gov/cotton/ars2.html

USDA Cotton Program
 www.ams.usda.gov/cotton/

USDA Cotton Briefing Room
 www.ers.usda.gov/Briefing/Cotton/
 www.ers.usda.gov/briefing/cotton/

USDA Key Topics -- Cotton
 www.ers.usda.gov/Topics/view.asp?T=101206
 www.ers.usda.gov/topics/view.asp?T=101206

Safety and Health Topics: Textiles (links to pages on cotton dust)
 www.osha-slc.gov/SLTC/textiles/

Southwestern Cotton Ginning Research Laboratory
 www.swcgrl.ars.usda.gov/
 www.swcgrl.ars.usda.gov/indextxt.htm
```

Figure 1: Example distillation topic with official qrels. Qrels give an overview of cotton sites in the .GOV corpus, therefore cotton activities in US Government.

```
<top>
<num> Number: NP151
<desc> Description:
ADA Enforcement
</top>

Answer: www.usdoj.gov/crt/ada/enforce.htm

--

<top>
<num> Number: NP161
<desc> Description:
TVA-Tennessee Valley Authority
</top>

Answer: www.tva.gov/
```

Figure 2: Example navigational topics with official qrel pages. The first is a 'named page' the second a 'homepage'.

2. Which techniques which have proven successful in homepage finding are also effective in named page finding?

3. Is it possible to identify homepage finding queries within a query stream?

Topics were numbered 151–450, with 150 homepage and 150 named page queries. For measures, we use the mean reciprocal rank of the first correct answer (MRR) and the proportion of queries where the correct answer appears in the top 10 (S@10). The number 10 was chosen for success rate calculation because search engines often provide 10 answers in the first page of search results.

3 Results

3.1 Topic Distillation Results

Across the 50 topics, 516 pages were judged relevant (average of 10.32 pages per query). Results for the best run submitted by each group are in Table 1. A full listing of runs is in Table 3.

Because there were only a few good answers for each query, system scores were low. This also seemed to reduce the stability of the results. The list of top 5 groups depends on how we sort Table 3. The top groups by R-Precision are in Table 1. Top groups by MAP were: CSIRO, Hummingbird, Neuchatel, UAmsterdam and UGlasgow. Top groups by P@10 were: Hummingbird, CSIRO, IBM Haifa, MSR Asia and UGlasgow.

Here we briefly summarize the information available about the experiments conducted by the top five groups (based on R-Precision).

CSIRO Documents scored via a linear combination of link indegree, anchortext propagation, URL Length and BM25. Linear combination (and BM25 parameters) were tuned using a home page finding query set (same tuning as navigational csiro03ki02). Stemming improved R-Precision by a further 0.0198.

Hummingbird Documents were given additional weight if their URL looked like a homepage URL, and also based on query word/phrase occurences in HTML markup such as title. There was no use of link counts or anchor text. Stemming had little effect.

UAmsterdam Used different representations and retrieval models. Okapi worked well on documents, titles and anchors. Language modelling worked very well on anchors and less well on documents and titles. Anchor text was important. Snowball stemming was used in all runs.

Copernic URL information was important (length and presence of query terms). Representations were each treated differently and included documents, extracted summaries, text with formatting, URL and title. First results were from a boolean AND query, followed by OR results. Porter stemming was used in all runs.

USunderland Used a novel document representation based on automatically assigned word senses as opposed to terms. The ranking algorithm consisted of a variation of Kleinberg's model of hubs and authorities in association with a number of vector space techniques including TF*IDF, and Cosine Similarity.

Based on information from these and other participants:

- Referring anchor text was important.
- Stemming was often helpful. Query expansion (blind feedback) was usually not necessary.
- URL information and link structure were helpful in several cases.
- Topic distillation was noted to bear some relationship to homepage finding, in terms of "what works".

Table 1: Best distillation run for each group, by R-Precision. The codes D, A, L indicate the use of document structure (D), Anchor text (A) and Link structure (L). Measures are R-Precision, mean average precision and precision at 10. (See appendix for a table of all runs.)

	R-Prec	MAP	P@10	Group	Run	D	A	L
1.	0.1636	0.1543	0.1240	csiro	csiro03td03	D	A	L
2.	0.1485	0.1387	0.1280	hummingbird	humTD03upl	D	_	_
3.	0.1432	0.1344	0.0980	uamsterdam	UAmsT03WtOk3	D	A	_
4.	0.1430	0.1325	0.0980	copernic	copTdRun5	D	_	_
5.	0.1407	0.1114	0.0940	usunderland	SBUNIQUE	D	_	L
6.	0.1391	0.1336	0.1140	uglasgow	uogtd4cahs	_	A	_
7.	0.1357	0.1371	0.0880	neuchatelu	UniNEtd4	D	A	L
8.	0.1354	0.1027	0.1160	microsoftasia	MSRA4002	D	A	L
9.	0.1262	0.1131	0.1060	tsinghuau	THUIRtd0305	D	A	L
10.	0.1173	0.1091	0.1220	ibmhaifa	JuruNoQDiff	D	A	L
11.	0.1096	0.0897	0.0920	umelbourne	MU03td01	D	A	_
12.	0.0918	0.0698	0.0920	meijiu	meijihilw1	D	_	_
13.	0.0906	0.0848	0.0760	vatech	VTtdgp5055	_	A	_
14.	0.0818	0.0799	0.0640	fub	fub03IneBBt	_	_	_
15.	0.0784	0.0818	0.0720	irit-sig	Merc1ti	_	_	_
16.	0.0769	0.0660	0.0440	kasetsartu	KUCONTENT	_	_	_
17.	0.0754	0.0728	0.0520	cas-ict	ICTWebTD12A	_	_	_
18.	0.0736	0.1016	0.0760	indianau	widittdb1	_	_	_
19.	0.0699	0.0896	0.0700	ajouu	ajouai0301	_	_	_
20.	0.0590	0.0691	0.0640	uillinoisuc	UIUC03W2s	_	_	_
21.	0.0395	0.0343	0.0280	lehighu	03wume206	_	A	L
22.	0.0281	0.0226	0.0320	umarylandbc	C2B	_	_	L
23.	0.0007	0.0001	0.0000	saarlandu	topics0	_	_	L

The small number of good answers for each query caused the measures to be much less stable than they have been in previous years. For P@10, systems should differ by at least 0.035 in absolute score for a

95% confidence in the difference, using the method of Voorhees and Buckley (SIGIR 2002). An absolute increase of 0.035 represents a 27.3% relative improvement over the top score, which is quite large. For R-Precision, the minimum absolute difference is 0.09. These differences represent quite a large swath of the full range of scores achieved.

The reason for this instability is as follows. Since more than half of the topics have less than 10 relevant, P@10 is always less than 1.0 for those topics, and so there's this overall degradation in the average due to those topics. The scores are also somewhat quantized, you don't see a continuous range of P@10. On the other hand, R-Precision has even more severe quantization, even though it averages better. Moreover, when there are less than 10 relevant documents, P@10 is an easier measure for systems to do well on than R-Precision, because they have more chances to get those documents.

3.2 Navigational Results

Judging involved identifying answers which appear at more than one URL. For example, the page `http://bernie.house.gov/imf/imf.asp` also appears at `http://www.bernie.house.gov/imf/imf.asp` so both were identified as correct answers for query 335. This process identified 19 queries with 2 URLs and 12 with more than 2.

Table 2: Best navigational run for each group, by mean reciprocal rank (MRR) of the first correct answer. The codes D, A, L indicate the use of document structure (D), Anchor text (A) and Link structure (L). Measures are MRR and the success rate at 10 as a percentage. (See appendix for a table of all runs.)

	MRR	S@10	Group	Run	D	A	L
1.	0.727	89.3	cmu	LmrEstUrl	D	A	_
2.	0.702	84.0	csiro	csiro03ki03	D	A	L
3.	0.688	84.7	neuchatelu	UniNEnp4	D	A	_
4.	0.665	87.0	iit	iit03sau	D	A	_
5.	0.651	84.3	microsoftasia	MSRANP1	D	A	_
6.	0.615	79.3	uglasgow	uogki2ca	_	A	_
7.	0.586	79.3	copernic	copNpRun1	D	_	_
8.	0.568	79.0	cas-ict	ICTWebKI12C	D	A	_
9.	0.561	81.0	tsinghuau	THUIRpf0301	_	A	_
10.	0.545	77.3	hummingbird	humNP03up	D	_	_
11.	0.530	75.0	umelbourne	MU03np4	D	A	_
12.	0.519	71.3	uamsterdam	UAmsT03WnLM3	D	A	_
13.	0.400	66.3	indianau	widitpff1	D	A	_
14.	0.374	59.0	vatech	VTnhpgp42	_	A	_
15.	0.350	53.7	rmit	RMITSEG3	D	_	_
16.	0.323	55.3	saarlandu	homepages0	_	_	_
17.	0.291	48.3	ajouu	ajouai0309	D	_	_
18.	0.120	16.3	ualaska	irttgrep	_	_	_
19.	0.067	9.3	lehighu	03wume298	_	A	_

The best run from each group is listed in Table 2. The full list of navigational runs is listed in Table 4.

Here we briefly summarize the information available about the experiments conducted by the top five groups (based on MRR).

CMU In order of MRR effectiveness, language models were formed from: in-link text, title text, full document text, image alternate text, modified size font text, meta keyword/description and a 3-character-gram URL. Best two runs used prior probabilities with URL classes from UTwente, trained on previous HP and NP tasks.

CSIRO Documents scored via a linear combination as described in the distillation section. Tunings were with a HP query set, NP query set and combined HP/NP query set. The NP tuning was most

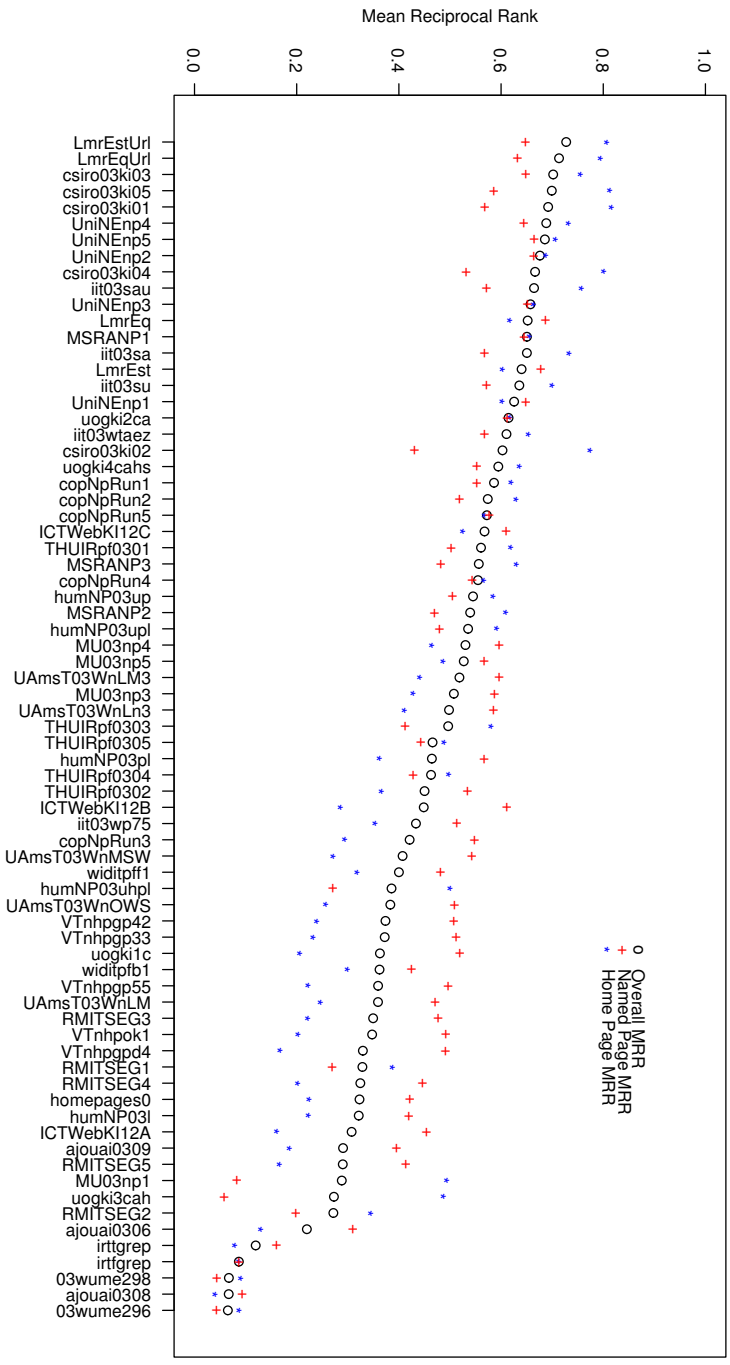

Figure 3: Breakdown of home page vs named page performance.

effective. Incorporating HP tuning was always harmful (in the combined tuning and in combinations via interleaving and CombSUM).

Neuchatel Document surrogates were 1) title with anchors from the current document, 2) title with anchors from referring documents and 3) various other segments of referring documents (title, h1, big). Scoring was with Okapi plus a proximity scoring function. Tuning was of the relative weight of surrogates, Okapi b parameter and the degree of proximity upweighting.

IIT Document surrogates were full text, title and anchors. Fusion was by CombMNZ with exponential z-score normalisation. A query task classification system was also employed, based on 32 words indicative of home page search such as 'home' or 'homepage'.

MSR Asia Document surrogates such as anchors and titles were employed, whose combination was tuned based on TREC-2002. Proximity information was used.

Based on information from these and other participants:

- It seems useful to consider different representations/surrogates based on document structure and including referring anchor text.
- Link structure gave mixed results.
- Stemming was not necessary for most participants.
- Several participants reported improvements based on proximity information, spans, phrases and fuzzy phrases.
- Few or no big wins from query classification (HP vs NP), but some promising attempts.

The scores for the navigational task are very stable. For the MRR measure on the full topic set, an absolute difference of 0.04 is enough to give 95% confidence. A 0.1 difference is reliable with as few as 100 topics.

There is only a little bit of a difference between named pages and homepages. If one uses the named page set alone, one needs to see an absolute difference between 0.06 and 0.07 for 95% confidence. For home pages, it's 0.06. If you pick 150 topics randomly from the larger set, it's 0.07, so the topic type does matter for the evaluation.

4 Conclusion

This year's topic distillation task was considered to be much more representative of real Web search than last year's. We also ran our first mixed query task, identifying approaches which work for home page and named page finding. Several groups attempted classification to differentiate between query types. More comprehensive mixed query tasks — possibly including topic distillation, topic relevance, service finding and navigational queries — seem to offer fertile ground for future evaluations.

Part II
Interactive Experiments

5 Introduction

For TREC 2003, the interactive track was a sub-track of the web track. The topic distillation, one of the web track tasks, has been selected as the interactive track task. The main motivation of the interactive sub-track was to investigate the role of human searchers in the topic distillation task.

6 Tasks

Eight search topics were selected from the topic set as used by the web topic distillation task. For each of these topics, a search scenario was provided in order to provide the participants a context for their search activity - it was not intended to boost the content available for searching.

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.

7 Search Systems

NIST provided the access to its server with the two versions of "Panoptic search engine". One version of the engine is optimized for the topic distillation task by balancing relevance and homepageness. The other, content-oriented, version is Okapi-based and retruns page in descending order of likely relevance. However, participants were free to use any appropriate search engine. To keep consistent with the automatic topic distillation task, all searches and browses were restricted to the .GOV collection.

8 Experimental Protocol

Participants were free to use any experimental protocol that suits their experimental purpose. However, the guidelines suggested an experiment design proposed by Rutgers group. Similar to the experiment designs in past interactive tracks, this design allows the comparison of two systems or system variants. This year's design divides the eight topics into two blocks, with varying order of topics within each block. This design requires a minimum 16 searchers, each searcher needs to search a block of four topics on one system and another block on another system.

9 Evaluation

The saved lists from each search session were gathered and sent to NIST for assessment. The assessment was based on four criteria: relevance, depth, coverage, and repetition. The assessors were asked to answer the following questions/statement on a five-point Likert scale.

Relevance: The page is relevant for the topic.

1 = Agree strongly, 2 = Agree slightly, 3 = Neutral, 4 = Disagree slightly, 5 = Disagree strongly

Depth: Is the page too broad, too narrow or at the right level of detail for the topic?

1 = Too broad, 2 = Bit broad, 3 = Right level, 4 = Bit narrow, 5 = Too narrow

Coverage: The set of saved entry points covers all the different aspects of the topic.

1 = Agree strongly, 2 = Agree slightly, 3 = Neutral, 4 = Disagree slightly, 5 = Disagree strongly

Repetition: How much repetition/overlap is there within the set of saved entry points?

1 = None, 2 = Minimal, 3 = Some, 4 = A lot of, 5 = Way too much

Instruments for the collection of searcher background and subjective evaluation of search systems were also provided and suggested by the Guidelines.

10 Overview of Results

Two groups, CSIRO and Rutgers, participated in this sub-track. Here is a brief description of their testing hypotheses and initial findings.

10.1 CSIRO

CSIRO investigated the effectiveness of a task tailored delivery method to assist searchers evaluate and thus select key resource pages as described by the topic distillation task.

In their baseline interface, they used the delivery interface from the Panoptic topic distillation engine which provides searchers with a ranked list of potential relevant key resource pages. In their testing interface, they designed a site summary interface and a sitemap interface to explicitly support searchers to judge whether a site is relevant and whether a page is at the right scope.

From their initial analysis, they found that their searchers preferred the testing interface and perceived that they fulfilled their task better by using the testing interface than the baseline interface. However, they didn't find any significant difference between the two interfaces on searcher's performance in terms of relevance, depth, coverage and repetition. By further examining searchers' behavior, they found that the interface for grouping documents into sites changed search behavior: searchers tend to assess a number of pages from the same site by reading their summary information before they selected a page to read further. Also in the post-system questionnaire, the searchers strongly stated that the grouping interface was useful for them to select an entry point to search. However, confounded by many other factors, it is not clear whether this behavior would be beneficial to the overall task.

Comparing the searchers' performance of the baseline interface with that of the corresponding automatic system, they found a significant improvement in terms of relevance, depth and precision. That indicates that engagement of searcher's effort has a positive effect on the system performance.

10.2 Rutgers

The Rutgers group investigated the role that the layout of search results plays in supporting human searchers executing topic distillation tasks. Success was measured in terms of accuracy and precision, operationalized as coverage and overlap, so the searcher was expected to find documents that provide information on as many distinct aspects of the assigned topic as possible, with as little overlap between them as possible. Their hypothesis was that using the structure of the domain and of the document corpus in order to organize the search output, would help identify aspects of the search topic in different sub-domains of the document collection, would reduce the searchers' cognitive load and would produce better results than the classic hit list. They tested this hypothesis by using two user interfaces for the Panoptic search engine, one with a simple list output, and the second with documents clustered based on common URL elements.

Their initial analysis shows that although it does not produce better coverage than the linear interface, the hierarchical interface seems to be conducive to less effort for the searcher: fewer iterations, shorter search sessions, fewer documents seen, selected and viewed. With regards to subjective measures, users perceived the hierarchical one as easier to use and better at supporting the topic distillation task. These results were not statistically significant. What was significant is that the subjects perceived the two systems equally easy to learn and that they preferred the hierarchical display.

One advantage of the structured output, as suggested by the objective measures and highlighted by the users' comments, is the support for investigating different sub-domains of a document collection and consequently different aspects of a topic. The searcher does not need to make a cognitive effort to separate the search results into sub-domain, so the layout makes the interaction easier and more pleasant and more accurately supports the searcher's judgment on task completion.

Appendix A: All non-interactive runs

Table 3: All distillation runs

	R-Prec	MAP	P@10	Group	Run	D	A	L
1.	0.1636	0.1543	0.1240	csiro	csiro03td03	D	A	L
2.	0.1485	0.1387	0.1280	hummingbird	humTD03upl	D	_	_
3.	0.1438	0.1354	0.1120	csiro	csiro03td01	D	A	L
4.	0.1432	0.1344	0.0980	uamsterdam	UAmsT03WtOk3	D	A	_
5.	0.1430	0.1325	0.0980	copernic	copTdRun5	D	_	_
6.	0.1407	0.1114	0.0940	usunderland	SBUNIQUE	D	_	L
7.	0.1391	0.1336	0.1140	uglasgow	uogtd4cahs	_	A	_
8.	0.1361	0.1284	0.1080	uglasgow	uogtd5cass	_	A	L
9.	0.1357	0.1371	0.0880	neuchatelu	UniNEtd4	D	A	L
10.	0.1354	0.1027	0.1160	microsoftasia	MSRA4002	D	A	L
11.	0.1333	0.1166	0.1020	usunderland	TBBASE	D	_	L
12.	0.1328	0.1198	0.1240	hummingbird	humTD03up	D	_	_
13.	0.1325	0.1273	0.1020	uglasgow	uogtd2ca	_	A	_
14.	0.1283	0.1259	0.1020	usunderland	SBBASE	D	_	L
15.	0.1282	0.1107	0.0960	copernic	copTdRun2	D	_	_
16.	0.1278	0.0948	0.0880	usunderland	TBUNIQUE	D	_	L
17.	0.1262	0.1131	0.1060	tsinghuau	THUIRtd0305	D	A	L
18.	0.1259	0.1187	0.0980	copernic	copTdRun3	D	_	_
19.	0.1217	0.1258	0.1080	csiro	csiro03td05	D	A	L
20.	0.1183	0.1180	0.0960	copernic	copTdRun1	D	_	_
21.	0.1173	0.1091	0.1220	ibmhaifa	JuruNoQDiff	D	A	L
22.	0.1162	0.1272	0.1060	csiro	csiro03td02	D	A	L
23.	0.1096	0.0897	0.0920	umelbourne	MU03td01	D	A	_
24.	0.1086	0.1127	0.0860	uamsterdam	UAmsT03WtOkC	_	_	_
25.	0.1079	0.1008	0.1220	ibmhaifa	JuruFull	D	A	L
26.	0.1078	0.0824	0.1100	microsoftasia	MSRA1002	D	A	_
27.	0.1069	0.0978	0.1020	hummingbird	humTD03uhpl	D	_	_
28.	0.1061	0.1022	0.1220	ibmhaifa	JuruNoCohes	D	A	L
29.	0.1060	0.0993	0.0660	copernic	copTdRun4	D	_	_
30.	0.1056	0.1019	0.0840	uamsterdam	UAmsT03WtLM3	D	A	_
31.	0.1053	0.1099	0.0820	uglasgow	uogtd3cas	_	A	_
32.	0.1052	0.0946	0.1140	microsoftasia	MSRA4003	D	A	L
33.	0.1046	0.1285	0.0980	neuchatelu	UniNEtd1	D	A	L
34.	0.1036	0.0764	0.0800	tsinghuau	THUIRtd0301	D	A	L
35.	0.1027	0.0699	0.0960	microsoftasia	MSRA1001	D	A	_
36.	0.1021	0.0902	0.0840	neuchatelu	UniNEtd5	D	_	L
37.	0.1016	0.0933	0.1040	microsoftasia	MSRA3	D	A	L
38.	0.1004	0.1041	0.0880	ibmhaifa	JuruNoAnchor	D	_	L
39.	0.0995	0.0982	0.0860	ibmhaifa	JuruNoSS	D	A	_
40.	0.0994	0.0763	0.0840	tsinghuau	THUIRtd0302	D	A	L
41.	0.0988	0.1004	0.1200	csiro	csiro03td04	D	_	L
42.	0.0918	0.0698	0.0920	meijiu	meijihilw1	D	_	_
43.	0.0906	0.0848	0.0760	vatech	VTtdgp5055	_	A	_
44.	0.0902	0.0652	0.1060	meijiu	meijihilw3	D	_	L
45.	0.0899	0.1003	0.0560	hummingbird	humTD03pl	D	_	_
46.	0.0823	0.0862	0.0760	uamsterdam	UAmsT03WtOkI	_	_	L
47.	0.0823	0.0898	0.0660	vatech	VTtdgp52	_	A	_
48.	0.0818	0.0799	0.0640	fub	fub03IneBBt	_	_	_
49.	0.0818	0.0818	0.0620	fub	fub03IneBMt	_	_	_
50.	0.0811	0.0864	0.0760	neuchatelu	UniNEtd2	D	A	_
51.	0.0811	0.0864	0.0760	neuchatelu	UniNEtd3	D	A	L
52.	0.0786	0.0646	0.0620	tsinghuau	THUIRtd0303	D	A	L

Table 3: All distillation runs (continued)

	R-Prec	MAP	P@10	Group	Run	D	A	L
53.	0.0786	0.0778	0.0620	fub	fub03InLBt	_	_	_
54.	0.0784	0.0818	0.0720	irit-sig	Merc1ti	_	_	_
55.	0.0783	0.0870	0.0760	irit-sig	Merc2tm	_	_	_
56.	0.0783	0.0775	0.0640	fub	fub03InBMt	_	_	_
57.	0.0769	0.0660	0.0440	kasetsartu	KUCONTENT	_	_	_
58.	0.0754	0.0728	0.0520	cas-ict	ICTWebTD12A	_	_	_
59.	0.0736	0.1016	0.0760	indianau	widittdb1	_	_	_
60.	0.0730	0.0886	0.0680	uglasgow	uogtd1c	_	_	_
61.	0.0728	0.0806	0.0580	umelbourne	MU03td05	D	_	_
62.	0.0716	0.0810	0.0580	fub	fub03InLBo1t	_	_	_
63.	0.0699	0.0896	0.0700	ajouu	ajouai0301	_	_	_
64.	0.0692	0.0558	0.0600	tsinghuau	THUIRtd0304	D	A	L
65.	0.0687	0.0486	0.0700	meijiu	meijihilw4	D	_	L
66.	0.0669	0.0845	0.0680	irit-sig	Merc2tp	_	_	_
67.	0.0655	0.0699	0.0540	vatech	VTtdgp33	_	A	_
68.	0.0652	0.0648	0.0580	ajouu	ajouai0305	D	_	_
69.	0.0648	0.0748	0.0620	vatech	VTtdok4	_	A	_
70.	0.0634	0.0687	0.0880	indianau	widittdb1r1	_	_	L
71.	0.0632	0.0639	0.0380	cas-ict	ICTWebTD12B	_	_	_
72.	0.0626	0.0787	0.0980	indianau	widittdf1r1	D	A	L
73.	0.0625	0.0647	0.0400	cas-ict	ICTWebTD12C	D	A	_
74.	0.0614	0.0486	0.0700	meijiu	meijihilw2	D	_	_
75.	0.0594	0.0733	0.0620	vatech	VTtdgp41	_	A	_
76.	0.0590	0.0691	0.0640	uillinoisuc	UIUC03W2s	_	_	_
77.	0.0590	0.0611	0.0580	uillinoisuc	UIUC03Wp	_	_	L
78.	0.0588	0.0773	0.0880	indianau	widittdf1r2	D	A	L
79.	0.0568	0.0631	0.0540	uillinoisuc	UIUC03Wu1	_	_	L
80.	0.0566	0.0627	0.0540	uillinoisuc	UIUC03Wb	_	_	_
81.	0.0559	0.0636	0.0400	umelbourne	MU03td04	D	A	_
82.	0.0553	0.0616	0.0540	uillinoisuc	UIUC03Wu2	_	_	L
83.	0.0537	0.0650	0.0660	umelbourne	MU03td03	D	A	_
84.	0.0523	0.0352	0.0620	meijiu	meijihilw5	D	_	L
85.	0.0433	0.0555	0.0400	irit-sig	Merc1td	_	_	_
86.	0.0395	0.0343	0.0280	lehighu	03wume206	_	A	L
87.	0.0391	0.0412	0.0280	uamsterdam	UAmsT03WtLMI	_	_	L
88.	0.0361	0.0512	0.0440	hummingbird	humTD03l	_	_	_
89.	0.0281	0.0226	0.0320	umarylandbc	C2B	_	_	L
90.	0.0230	0.0222	0.0200	umarylandbc	C2A	_	_	L
91.	0.0204	0.0225	0.0180	lehighu	03wume359	_	A	L
92.	0.0181	0.0250	0.0160	ajouu	ajouai0302	D	_	L
93.	0.0007	0.0001	0.0000	saarlandu	topics0	_	_	L

Table 4: All navigational runs

	MRR	S@10	Group	Run	D	A	L
1.	0.727	89.3	cmu	LmrEstUrl	D	A	_
2.	0.713	88.0	cmu	LmrEqUrl	D	A	_
3.	0.702	84.0	csiro	csiro03ki03	D	A	L
4.	0.699	81.0	csiro	csiro03ki05	D	A	L
5.	0.692	83.7	csiro	csiro03ki01	D	A	L
6.	0.688	84.7	neuchatelu	UniNEnp4	D	A	_
7.	0.686	84.7	neuchatelu	UniNEnp5	D	A	_
8.	0.676	84.0	neuchatelu	UniNEnp2	D	A	_

Table 4: All navigational runs

	MRR	S@10	Group	Run	D	A	L
9.	0.667	86.3	csiro	csiro03ki04	D	A	L
10.	0.665	87.0	iit	iit03sau	D	A	_
11.	0.658	83.7	neuchatelu	UniNEnp3	D	A	_
12.	0.652	83.3	cmu	LmrEq	D	A	_
13.	0.651	84.3	microsoftasia	MSRANP1	D	A	_
14.	0.651	86.7	iit	iit03sa	D	A	_
15.	0.640	83.3	cmu	LmrEst	D	A	_
16.	0.636	85.7	iit	iit03su	D	A	_
17.	0.626	82.3	neuchatelu	UniNEnp1	D	A	_
18.	0.615	79.3	uglasgow	uogki2ca	_	A	_
19.	0.611	84.0	iit	iit03wtaez	D	A	_
20.	0.603	77.7	csiro	csiro03ki02	D	A	L
21.	0.595	75.7	uglasgow	uogki4cahs	_	A	_
22.	0.586	79.3	copernic	copNpRun1	D	_	_
23.	0.574	77.0	copernic	copNpRun2	D	_	_
24.	0.572	75.7	copernic	copNpRun5	D	_	_
25.	0.568	79.0	cas-ict	ICTWebKI12C	D	A	_
26.	0.561	81.0	tsinghuau	THUIRpf0301	_	A	_
27.	0.556	72.7	microsoftasia	MSRANP3	D	A	_
28.	0.555	74.7	copernic	copNpRun4	D	_	_
29.	0.545	77.3	hummingbird	humNP03up	D	_	_
30.	0.540	71.3	microsoftasia	MSRANP2	D	A	_
31.	0.535	77.7	hummingbird	humNP03upl	D	_	_
32.	0.530	75.0	umelbourne	MU03np4	D	A	_
33.	0.527	76.0	umelbourne	MU03np5	D	A	_
34.	0.519	71.3	uamsterdam	UAmsT03WnLM3	D	A	_
35.	0.508	76.7	umelbourne	MU03np3	D	A	_
36.	0.498	72.7	uamsterdam	UAmsT03WnLn3	D	A	_
37.	0.496	64.3	tsinghuau	THUIRpf0303	_	A	_
38.	0.466	63.7	tsinghuau	THUIRpf0305	_	A	_
39.	0.465	68.3	hummingbird	humNP03pl	D	_	_
40.	0.463	62.7	tsinghuau	THUIRpf0304	_	A	_
41.	0.450	75.3	tsinghuau	THUIRpf0302	_	A	_
42.	0.449	65.7	cas-ict	ICTWebKI12B	D	A	_
43.	0.433	67.0	iit	iit03wp75	_	_	_
44.	0.421	61.3	copernic	copNpRun3	D	_	_
45.	0.407	63.0	uamsterdam	UAmsT03WnMSW	_	_	_
46.	0.400	66.3	indianau	widitpff1	D	A	_
47.	0.386	56.7	hummingbird	humNP03uhpl	D	_	_
48.	0.383	59.3	uamsterdam	UAmsT03WnOWS	_	_	_
49.	0.374	59.0	vatech	VTnhpgp42	_	A	_
50.	0.372	59.0	vatech	VTnhpgp33	_	A	_
51.	0.363	55.7	uglasgow	uogki1c	_	_	_
52.	0.362	60.0	indianau	widitpfb1	_	_	_
53.	0.359	56.7	uamsterdam	UAmsT03WnLM	_	_	_
54.	0.359	57.7	vatech	VTnhpgp55	_	A	_
55.	0.350	53.7	rmit	RMITSEG3	D	_	_
56.	0.348	55.7	vatech	VTnhpok1	_	A	_
57.	0.330	51.3	vatech	VTnhpgpd4	_	A	_
58.	0.329	55.3	rmit	RMITSEG1	D	_	_
59.	0.325	54.0	rmit	RMITSEG4	D	_	_
60.	0.323	55.3	saarlandu	homepages0	_	_	_
61.	0.321	54.3	hummingbird	humNP03l	_	_	_
62.	0.308	54.0	cas-ict	ICTWebKI12A	_	_	_
63.	0.291	48.3	ajouu	ajouai0309	D	_	_

Table 4: All navigational runs

	MRR	S@10	Group	Run	D	A	L
64.	0.290	48.0	rmit	RMITSEG5	D	_	_
65.	0.288	40.0	umelbourne	MU03np1	D	A	_
66.	0.273	39.0	uglasgow	uogki3cah	_	A	_
67.	0.272	43.7	rmit	RMITSEG2	D	_	_
68.	0.220	41.3	ajouu	ajouai0306	D	_	_
69.	0.120	16.3	ualaska	irttgrep	_	_	_
70.	0.087	17.3	ualaska	irtfgrep	_	_	_
71.	0.067	9.3	lehighu	03wume298	_	A	_
72.	0.067	11.7	ajouu	ajouai0308	D	_	L
73.	0.065	8.7	lehighu	03wume296	_	A	_