Overview of the TREC-2002 Web Track

Nick Craswell and David Hawking
CSIRO Mathematical and Information Sciences,
Canberra, Australia
{Nick.Craswell,David.Hawking}@csiro.au

April 16, 2003

Abstract

The TREC-2002 Web Track moved away from non-Web relevance ranking and towards Web-specific tasks on a 1.25 million page crawl ".GOV". The topic distillation task involved finding pages which were relevant, but also had characteristics which would make them desirable inclusions in a distilled list of key pages. The named page task is a variant of last year's homepage finding task. The task is to find a particular page, but in this year's task the page need not be a home page.

1 Introduction

The TREC-2002 Web Track activities centred on two tasks: A Topic Distillation Task and a Named Page Finding Task. Both made use of an 18 gigabyte, 1.25 million document 2002 partial crawl of the .gov domain, distributed on CD-ROM as the .GOV collection.

2 Guidelines

2.1 This Year's Aims

1. To begin work with a new (early 2002) crawl of an important Web domain (.gov). Past TREC Web experiments used data from 1997.
2. To formulate Web-specific search tasks, which are representative of common Web search activities, leading to new evaluation methods and new effective Web retrieval algorithms.
3. To conduct topic distillation experiments, in order to understand the selectivity required to generate a short top-N list, even when a very large set of on-topic documents are available.
4. To conduct named page experiments, to find if there are particular forms of ranking evidence which help us to find specific Web documents (last year's experiments found that URL type and anchor text were useful for finding homepage documents).
5. To make available the first set of reusable relevance judgments for the new .GOV test collection.

2.2 Dataset

The .GOV corpus is a crawl of Web sites in the .gov domain from early 2002. That makes it 5 years newer than previous TREC Web collections, all of which were based on a 1997 Internet Archive crawl. Although we hope that the most useful of the Web search techniques would work on 1997 crawls as well as 2002, it is also highly desirable to have a dataset representative of the current Web. Some properties of .GOV are listed in Table 1.

Table 1: Salient properties of the .GOV corpus. (Mime types as reported by the servers.)

Number of pages	1,247,753
Number of pages by mime type:	
text/html	1,053,110
application/pdf	131,333
text/plain	43,753
application/msword	13,842
application/postscript	5,673
other (containing text)	42
Average page size	15.2 kB
Number of hostnames	7,794
Total number of links	11,164,829
Number of cross-host links	2,470,109
Average cross-host links per host	317

The crawl included binary and text mime types, and was stopped after 1 million HTML pages. The HTML and text, plus extracted text of other document types, gives a total of 1.25 million documents. Total data size was 35 gigabytes, which we considered too great an increase in corpus size (over WT10g), so a 100 kilobyte cutoff was applied to all documents, reducing the total size to 18 gigabytes.

The .GOV dataset is distributed by CSIRO [3]. Note that the standard distribution includes the HTML documents plus text extracted from other formats such as PDF. The original PDFs and other binary files such as images were collected and are potentially available. However, the full crawl including binaries is 67 gigabytes almost four times the size of the collection as distributed (and it would not compress as well).

Docids are 14 characters and of the form G09-04-2395783, meaning that this document is in the bundle G00/04.gz at byte offset 2395783. All .GOV documents can be located in this way via their docid. The collection was distributed on seven CDs, with an eighth containing tables of URLtoID, links, duplicates and redirects. The URLtoID table lists all valid .GOV docids and their corresponding URLs. The link table is useful for link-based ranking experiments, and a potentially more complete picture of link structure could be built in conjunction with the provided duplicate and redirect tables. These give information on link target URLs visited by the crawler but not included in .GOV because they contained duplicate content or forwarded users to another URL.

The .GOV corpus has fewer documents than WT10g, but has a much larger average document size (15k vs 7k), reflecting changes in Web authoring over the space of five years (perhaps the prevalence of navigation bars and scripting in more recent pages). Compared to WT10g, .GOV also has the strictest file-type checking of any Web collection so far, leading to very few binaries in the corpus.

We chose to crawl .gov for several reasons. It is a commercially interesting domain, meaning that important services are provided based on precisely this sort of crawl. It is also a crawl of manageable size, in that it can be distributed on CD and is within the data size limitations of most TREC systems. By contrast the crawl of a large search engine would be perhaps 30 terabytes, well beyond the bounds of manageability using current technologies (the 100 gigabyte VLC2 is still considered large relative to the storage media and systems available to researchers). Luckily, many smaller crawls such as .GOV are conceivable, which are of manageable size and of significant research and commercial interest. The .GOV crawl is also of a size which allows sufficiently complete relevance judgments.

2.3 Topic Distillation Task

Topics: 551-600 Example:

<top>

```
<num> Number: 600
<title> highway safety
<desc> Description:
Find documents related to improving highway safety in the U.S.
<narr> Narrative:
Relevant documents include those related to the improvement of safety
of all vehicles driven on highways, including cars, trucks, vans, and
tractor trailers. Ways to reduce accidents through legislation,
vehicle checks, and drivers education programs are all relevant.
</top>
```

The premise of the topic distillation task is that some quality, in addition to relevance, is desirable in Web search results lists. This quality has been called authority, quality, definitiveness and many other names in previous studies [4, 1, 2]. Assuming it exists, systems will have to find evidence which indicates its presence, and strike a balance between relevance and "quality" in search algorithms. This balancing act is analogous to the balance between newness and relevance when searching a news archive: it is desirable to return documents which are both relevant and new (if any).

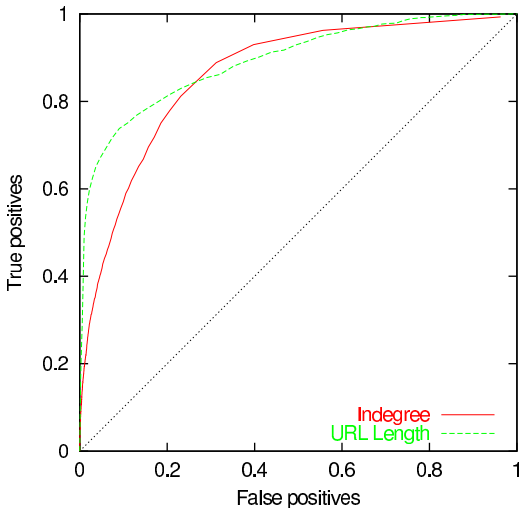

Figure 1: Query-independent properties of .GOV pages predict which will be listed in Web directories.

In our non-TREC research, we have found some evidence that query-independent evidence can indicate desirability, based on analysis of hand-made URL lists. We sorted all .GOV URLs in reverse URL length order and link indegree order. Then we found examples of "quality URLs" in .GOV, by identifying those which are hand-listed in the Yahoo! and DMOZ Web directories. Web directories are an alternative Web search technology, where within a category hierarchy, each category has a list of URLs created by a human editor. Figure 1 shows that the URL and link orderings were good predictors of hand-listing. If they predict hand-listing, then these characteristics might also be good predictors of which search results would be preferred by Web search users (who, after all, are also the audience of the Web directories).

In the topic distillation task we evaluate systems in terms of their ability to return relevant "key pages". A key page is one which the relevance assessor would find worthy of including in a short list of important URLs (the sort of choice made by Web directory editors). The "relevant key pages" found by assessors should thus be relevant and posses that special quality which makes pages worthy of inclusion in a short list. We did not go further than this in defining what makes a page list-worthy, since it has not been agreed in the research community what the definition is (quality, authority, definitiveness etc), and we did not want to bias assessments. The main measure is precision at 10.

2.4 Named Page Finding Task

Topics: NP1-NP150 Example:

```
<top>
<num> Number: NP1
<desc> Description:
America's Century Farms
</top>
```

The objective in the named page finding task is to find a particular Web page in .GOV, given a query which describes it by name. For example, the query "America's Century Farms" might lead to a particular .GOV Web page describing farms that have remained in one family's hands for over 100 years. The assessment task was simply to identify any duplicate URLs for the named page, since a page can appear at more than one URL. The main measure was the reciprocal rank of the first correct answer.

2.5 Indexing Restrictions

There were none. Participants were permitted to index all of each document or exclude certain fields as they wished.

2.6 Submissions and Judgments

Runs were received from a total of 23 groups: ajou, chinese_academy, city-pliers, cmu_lti, csiro, cuny, dgic_stokoe, fudan, glasgow, hummingbird, ibm-haifa, iit, illinois_chicago, irit, kasetsart, lit_singapore, neuchatel, tsinghua, umbc-cost, umelbourne, uva, waterloo, yonsei.

2.7 Topic Distillation Task

Seventy-one official runs were submitted from seventeen participating groups. The number of pages judged was 56,650 of which 1574 were judged to meet the criteria. Figure 2 shows the distribution of numbers of key resources found per topic. For a few topics the number of such resources is very much higher than expected.

While pages hand-listed in Web directories tended to have short URLs and high indegree (Figure 1), key resources from this year's track did not show such tendencies as strongly (Figure 3).

2.8 Named Page Finding Task

Seventy official runs were submitted from eighteen participating groups.

Only one correct answer was identified for most of the 150 topics, but there were three correct answers to two topics (9 and 145) and two for 16 topics (1, 8, 14, 24, 26, 50, 51, 63, 66, 67, 68, 85, 89, 128, 138, and 146).

3 Results

3.1 Topic Distillation Task

Full official results for the Topic Distillation task are reported in Appendix 1. Results on a per-group basis are presented in Table 2. Here we briefly summarize the information available about the experiments conducted by the top five groups

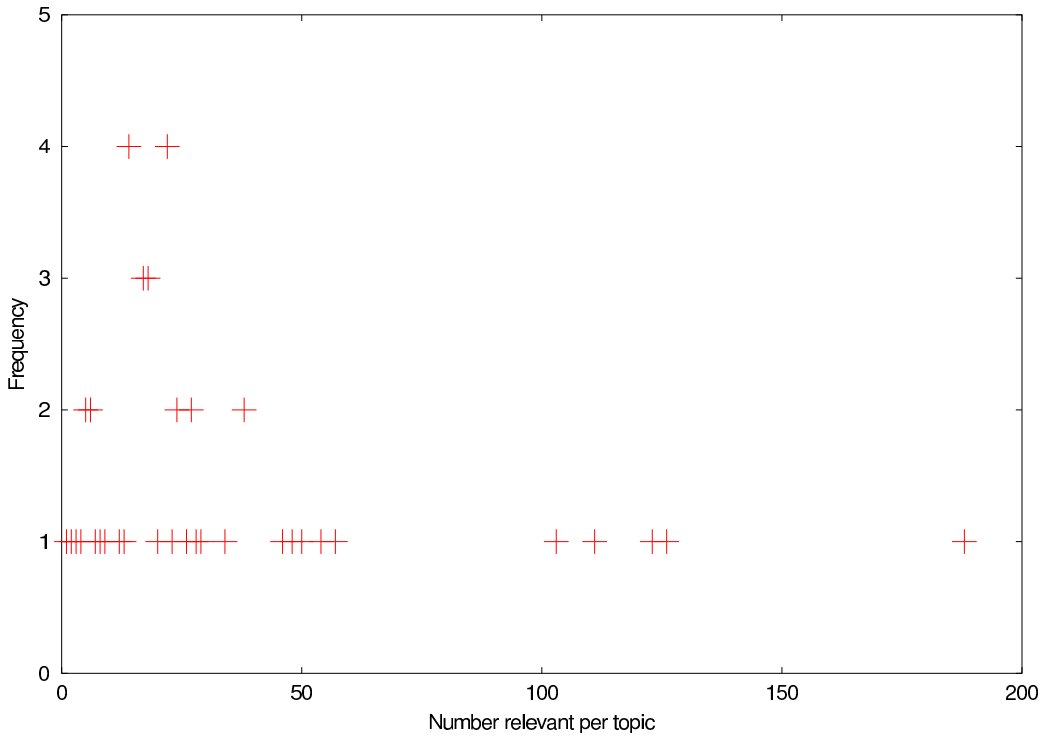

Figure 2: Topic distillation task: Number of key resources per topic.

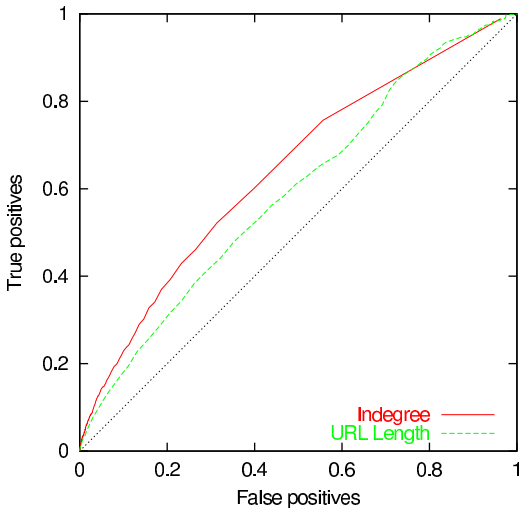

Figure 3: Query-independent properties which were good predictors in Figure 1 are less useful when predicting this year's key pages.

TsingHua University TsingHua used Okapi with stemming and Fox stoplist but no query expansion or feedback.

They explored:

1. techniques based on link structure and link text, especially the use of out-degree to find key resources;

Table 2: P@10 results for the best topic distillation run submitted by each participating group. The codes D, A, L indicate the use of document structure (D), Anchor text (A) and Link structure (L).

Rank	P@10	Group	Best Run	Run type	D?	A?	L?	#Runs
1.	0.2510	tsinghua	thutd5	realistic	D	A	_	5
2.	0.2408	city-pliers	pltr02wt2	realistic	_	_	_	2
3.	0.2306	chinese_academy	icttd1	realistic	_	_	_	2
4.	0.2286	ibm-haifa	ibmhaifapr	realistic	D	A	L	4
5.	0.2224	glasgow	uog05tad	realistic	D	A	L	2
6.	0.2163	irit	mercah	realistic	_	_	_	2
7.	0.1959	neuchatel	uninedi5	realistic	D	A	_	5
8.	0.1939	fudan	fduwt11t1	realistic	D	_	L	3
9.	0.1939	umelbourne	mu525	realistic	_	_	_	5
10.	0.1755	uva	uamst02wtt	realistic	_	_	_	5
11.	0.1510	yonsei	yedi01	realistic	D	A	L	1
12.	0.1143	umbc-cost	carrot2a	realistic	_	_	_	1
13.	0.1082	cuny	pirc2wd2	realistic	D	A	L	2
14.	0.1041	illinois_chicago	uic0104	realistic	_	_	L	2
15.	0.1000	csiro	csiro02td1	realistic	_	_	L	1
16.	0.0714	dgic_stokoe	tdwsdtfidf	realistic	_	_	_	2
17.	0.0571	ajou	ajouai0210	realistic	_	_	L	4

2. the roles of different HTML fields in ranking content;
3. post-processing of retrieval results, namely a site uniting approach
4. A genetic algorithm based dynamic parameter learning approach.

They found that anchor text was useful but out-degree was not. They also found that site uniting methods which worked well on the small number of training examples improved average precision but not P@10. Parameter settings learned on past Web tasks did not improve performance this year.

City University, London It is significant that the second best performing run (`pltr02wt2` from City University, London) was a straightforward content retrieval run based on Okapi BM25 (with non-default parameter for parameter b and stemming but no relevance feedback.

Chinese Academy No details available.

IBM Haifa Query expansion via lexical affinities. Knowledge Agents and Knowledge Bases incorporating content scores, anchor text, Kleinberg Hub and Authority scores and SALSA scores. Site compression. Title filtering - eliminate documents which have no query word in their title (beneficial). Duplicate elimination based on textual similarity (harmful).

Glasgow University Experimentation focused on:

1. A probabilistic framework for combining link and content, called the Absorbing Model, based on Markov chains and applicable in either static or dynamic forms;
2. A spreading activation method (either query independent or query dependent) for detecting site entry points;
3. Anchor text.
4. A genetic algorithm based dynamic parameter learning approach.

They found that body only indexing and link analysis without anchors work well while spreading activation on sites was equivocal and query expansion and PageRank were detrimental.

IBM Haifa identified a scoring problem in that no penalty was applied to runs which included multiple duplicates or near duplicates.

3.2 Named Page Finding Task

Table 3: MRR results for the best named page finding run submitted by each participating group. The codes D, A, L indicate the use of document structure (D), Anchor text (A) and Link structure (L).

Rank	MRR	Group	Run	Run Type	D?	A?	L?
1	0.719	tsinghua	thunp3	realistic	D	A	_
2	0.676	cmu_lti	lmralleq	realistic	D	A	_
3	0.671	yonsei	yenp01	realistic	D	A	L
4	0.654	glasgow	uog07cta	realistic	D	A	_
5	0.636	neuchatel	uninenp1	realistic	D	A	_
6	0.626	hummingbird	hum02pd	realistic	D	_	_
7	0.613	chinese_academy	ictnp6	realistic	D	A	_
8	0.587	iit	iit02b	realistic	_	_	_
9	0.578	lit_singapore	litlink	realistic	D	A	L
10	0.576	umelbourne	mu106	realistic	D	A	_
11	0.573	csiro	csiro02np01	realistic	_	_	_
12	0.564	illinois_chicago	uicnp03	realistic	_	_	_
13	0.535	waterloo	uwmtbw2	realistic	_	A	L
14	0.432	uva	uamst02wntla	realistic	_	A	_
15	0.418	city-pliers	pltr02wt9	realistic	_	_	_
16	0.263	cuny	pirc2wnp1	realistic	D	A	_
17	0.132	ajou	ajouai0204	realistic	D	_	_
18	0.010	kasetsart	kuhpf0201	realistic	_	A	_

Full official results for the Named Page Finding task are reported in Appendix 2. Results on a per-group basis are presented in Table 3. Here we briefly summarize the information available about the experiments conducted by the top five groups

TsingHua University They built a collection of surrogate documents comprising keywords, titles and incoming anchor text. Ranks obtained with these surrogates were combined with ranks from the original documents, using $S' = a*1/rank1+(1-a)*1/rank2$. (Note that the original collection was divided into two sub-collections: html and non-html and the results merged using a novel procedure (see the TsingHua paper for details). The combined score outperformed the original score which in turn outperformed the surrogate score.

CMU LTI Their basic model was a generative language model, where the language model for the document was a linear interpolation of several language models (title, in-link text, full text, meta tag text, image alt text, url text, large fonts). Using document structure in this way did improve performance over just using a simple language model.

They were unable to find useful prior probabilities for this task, either on training data created locally or on the test data. They tried in-link count, document length, document file type, and url length.

Yonsei No details available.

Glasgow University Anchor text proved to be more useful than link analysis, significantly improving results.

They found that body only indexing and link analysis without anchors work well while spreading activation on sites was equivocal and query expansion and PageRank were detrimental.

U. Neuchatel A second representation of each document in the .GOV collection was created, comprising the documents title and all its incoming anchor text. Okapi scores were computed for both representations and linearly combined $\alpha S_{content} + (1-\alpha) S_{anchortitle}$ (without normalisation). The best results were obtained with $\alpha = 0.6$.

4 Conclusions

The .GOV corpus provided an interesting and realistic dataset for the purposes of the track. No significant problems were reported in working with it.

The Named Page Finding task was an interesting variant on earlier Home Page Finding evaluations. Unsurprisingly, URL-type analyses did not bring improvement in performance. However, several leading participants reported an improvement in performance by adding anchor text and structural information to a content-only run. In 2003, it is anticipated that a mixed Home Page / Named Page task might prove interesting.

The Topic Distillation task proved difficult to explain to both participants and assessors and there was considerable disparity between the interpretations of these two groups. It is not clear what, if any, conclusions can be drawn at this stage. The task is worth repeating in 2003 but more explanatory effort is needed.

Acknowledgements

We are very grateful to Charlie Clarke of the University of Waterloo for providing the crawler software used to collect the .GOV data and for supervising the crawl. We are also grateful to Ed Fox of Virginia Tech for providing the machine and network connection to support the crawl and to Ian Soboroff of NIST for facilitating the crawl and also for his invaluable assistance in organising the track.

References

[1] K. Bharat and M. Henzinger. Improved algorithms for topic distillation in a hyperlinked environment. In *Proceedings of ACM SIGIR'98*, pages 104–111, 1998.

[2] S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. In *Proceedings of WWW7*, pages 107–117, 1998. `http://www7.scu.edu.au/programme/fullpapers/1921/com1921.htm`.

[3] CSIRO. TREC Web Tracks home page. `www.ted.cmis.csiro.au/TRECWeb/`.

[4] Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.

Appendix 1 - Topic Distillation runs

P@10 results for all topic distillation run submitted. The codes D, A, L indicate the use of document structure (D), Anchor text (A) and Link structure (L)

Rank	P@10	Group	Run	Run type	D?	A?	L?	#Runs
1.	0.2510	tsinghua	thutd5	realistic	D	A	-	5
2.	0.2408	city-pliers	pltr02wt2	realistic	-	-	-	2
3.	0.2306	tsinghua	thutd3	realistic	D	A	-	3
4.	0.2306	chinese_academy	icttd1	realistic	-	-	-	2
5.	0.2286	ibm-haifa	ibmhaifapr	realistic	D	A	L	4
6.	0.2245	tsinghua	thutd2	realistic	D	A	-	2
7.	0.2224	glasgow	uog05tad	realistic	D	A	L	2
8.	0.2163	irit	mercah	realistic	-	-	-	2
9.	0.2143	tsinghua	thutd4	realistic	D	A	-	4
10.	0.2122	ibm-haifa	ibmhaifat10	realistic	D	A	L	3
11.	0.2082	glasgow	uog04cta2dqh	realistic	D	A	L	4
12.	0.2061	ibm-haifa	ibmhaifat10d	realistic	D	A	L	3
13.	0.2000	city-pliers	pltr02wt1	realistic	-	-	-	1
14.	0.2000	city-pliers	pltr02wt4	realistic	-	-	-	4
15.	0.1980	tsinghua	thutd1	realistic	D	A	L	1
16.	0.1959	neuchatel	uninedi5	realistic	D	A	-	5
17.	0.1939	ibm-haifa	ibmhaifaap	realistic	D	A	L	2
18.	0.1939	fudan	fduwt11t1	realistic	D	-	L	3
19.	0.1939	fudan	fduwt11o1	realistic	D	-	L	1
20.	0.1939	glasgow	uog03ctadqh	realistic	D	A	L	1
21.	0.1939	umelbourne	mu525	realistic	-	-	-	5
22.	0.1939	fudan	fduwt11t2	realistic	D	A	L	2
23.	0.1898	ibm-haifa	ibmhaifabase	realistic	D	A	L	1
24.	0.1857	umelbourne	mu111	realistic	D	A	-	1
25.	0.1755	city-pliers	pltr02wt3	realistic	-	-	-	3
26.	0.1755	uva	uamst02wtt	realistic	-	-	-	5
27.	0.1755	chinese_academy	icttd2	realistic	-	-	-	1
28.	0.1714	fudan	fduwt11o2	realistic	D	A	L	4
29.	0.1694	umelbourne	mu624	realistic	-	-	-	2
30.	0.1673	chinese_academy	icttd3	realistic	-	-	-	3
31.	0.1510	fudan	fduwt11b0	realistic	-	-	-	5
32.	0.1510	yonsei	yedi01	realistic	D	A	L	1
33.	0.1469	yonsei	yedi01no	realistic	D	A	L	2
34.	0.1429	irit	mercure	realistic	-	-	L	3
35.	0.1429	neuchatel	uninedi4	realistic	D	A	-	4
36.	0.1306	glasgow	uog01ctaialh	realistic	D	A	L	5
37.	0.1163	umelbourne	mu313	realistic	D	A	-	3
38.	0.1143	umbc-cost	carrot2a	realistic	-	-	-	1
39.	0.1143	glasgow	uog02ctadh	realistic	D	A	L	3
40.	0.1082	umelbourne	mu212	realistic	D	A	-	1
41.	0.1082	irit	mercurelynx	realistic	-	-	L	1
42.	0.1082	cuny	pirc2wd2	realistic	D	A	L	2
43.	0.1041	illinois_chicago	uic0104	realistic	-	-	L	2
44.	0.1000	csiro	csiro02td1	realistic	-	-	L	1
45.	0.1000	illinois_chicago	uic0101	realistic	-	-	L	4
46.	0.1000	uva	uamst02wta	realistic	-	A	-	4
47.	0.0939	csiro	csiro02td5	realistic	-	-	L	5
48.	0.0898	illinois_chicago	uic0103	realistic	-	-	L	3
49.	0.0878	city-pliers	pltr02wt5	exploratory	-	-	-	5
50.	0.0837	neuchatel	uninedi1	realistic	D	A	L	1
51.	0.0816	cuny	pirc2wd1	realistic	D	A	L	1
52.	0.0796	illinois_chicago	uic0102	realistic	-	-	L	1
53.	0.0776	neuchatel	uninedi3	exploratory	D	A	L	2
54.	0.0714	csiro	csiro02td2	realistic	-	A	L	2
55.	0.0714	dgic_stokoe	tdwsdtfidf	realistic	-	-	-	2
56.	0.0714	umbc-cost	carrot2c	realistic	-	-	L	4
57.	0.0673	uva	uamst02wttri	realistic	-	-	L	3
58.	0.0653	uva	uamst02wtacs	realistic	-	A	-	2
59.	0.0653	dgic_stokoe	tdtfidf	realistic	-	-	-	1
60.	0.0633	uva	uamst02wtari	realistic	-	A	L	1
61.	0.0571	ajou	ajouai0210	realistic	-	-	L	4
62.	0.0551	umbc-cost	carrot2d	realistic	-	-	L	2
63.	0.0551	ajou	ajouai0206	realistic	-	-	L	2
64.	0.0551	umbc-cost	carrot2b	realistic	-	-	-	3
65.	0.0531	ajou	ajouai0207	realistic	-	-	L	1
66.	0.0347	ajou	ajouai0208	realistic	-	-	L	3
67.	0.0327	neuchatel	uninedi2	realistic	D	A	L	3
68.	0.0245	ajou	ajouai0209	realistic	-	-	L	5
69.	0.0184	csiro	csiro02td4	realistic	-	A	L	4
70.	0.0184	umbc-cost	carrot2e	realistic	-	-	L	4
71.	0.0184	csiro	csiro02td3	realistic	-	-	L	3

Appendix 2 - Named Page Finding runs

MRR results for all named page finding run submitted. The codes D, A, L indicate the use of document structure (D), Anchor text (A) and Link structure (L).

Rank	MRR	Group	Run	Run Type	D?	A?	L?
1	0.719	tsinghua	thunp3	realistic	D	A	-
2	0.717	tsinghua	thunp5	realistic	D	A	-
3	0.690	tsinghua	thunp1	realistic	D	A	-
4	0.687	tsinghua	thunp4	realistic	D	A	-
5	0.676	cmu_lti	lmralleq	realistic	D	A	-
6	0.671	yonsei	yenp01	realistic	D	A	L
7	0.667	cmu_lti	lmrallest	realistic	D	A	-
8	0.654	glasgow	uog07cta	realistic	D	A	-
9	0.651	glasgow	uog10ctad	realistic	D	A	L
10	0.643	glasgow	uog09cta2	realistic	D	A	-
11	0.636	neuchatel	uninenp1	realistic	D	A	-
12	0.626	hummingbird	hum02pd	realistic	D	-	-
13	0.625	neuchatel	uninenp3	realistic	D	A	-
14	0.616	neuchatel	uninenp2	realistic	D	A	-
15	0.613	chinese_academy	ictnp7	realistic	D	A	-
16	0.613	chinese_academy	ictnp6	realistic	D	A	-
17	0.611	cmu_lti	lmrnostruct	realistic	-	A	-
18	0.589	cmu_lti	lmrsmall	realistic	D	A	-
19	0.587	iit	iit02b	realistic	-	-	-
20	0.580	iit	iit02tfa	realistic	D	A	-
21	0.578	lit_singapore	litlink	realistic	D	A	L
22	0.576	umelbourne	mu106	realistic	D	A	-
23	0.576	iit	iit02tf	realistic	D	-	-
24	0.573	csiro	csiro02np01	realistic	-	-	-
25	0.568	cmu_lti	lmrdocstruct	realistic	D	-	-
26	0.564	illinois_chicago	uicnp03	realistic	-	-	-
27	0.559	chinese_academy	ictnp2	realistic	D	A	-
28	0.557	chinese_academy	ictnp3	realistic	D	A	-
29	0.555	chinese_academy	ictnp4	realistic	D	A	-
30	0.552	glasgow	uog06c	realistic	-	-	-
31	0.550	illinois_chicago	uicnp02	realistic	-	-	-
32	0.538	hummingbird	hum02upd	realistic	D	-	-
33	0.535	waterloo	uwmtbw2	realistic	-	A	L
34	0.530	tsinghua	thunp2	realistic	D	A	-
35	0.527	hummingbird	hum02up	realistic	D	-	-
36	0.524	umelbourne	mu609	realistic	D	A	-
37	0.524	umelbourne	mu208	realistic	D	A	-
38	0.516	glasgow	uog08ctap	realistic	D	A	-
39	0.509	waterloo	uwmtbw0	realistic	-	-	-
40	0.504	neuchatel	uninenp4	realistic	D	A	-
41	0.495	illinois_chicago	uicnp01	realistic	-	-	L
42	0.456	hummingbird	hum02ud	realistic	-	-	-
43	0.432	uva	uamst02wntla	realistic	-	A	-
44	0.427	lit_singapore	littext	realistic	-	-	-
45	0.425	uva	uamst02wntl	realistic	-	-	-
46	0.418	city-pliers	pltr02wt9	realistic	-	-	-
47	0.416	csiro	csiro02np03	realistic	D	-	-
48	0.416	city-pliers	pltr02wt8	realistic	-	-	-
49	0.414	city-pliers	pltr02wt7	realistic	-	-	-
50	0.402	umelbourne	mu80a	realistic	-	-	-
51	0.367	uva	uamst02wntma	realistic	-	A	-
52	0.337	hummingbird	hum02uhp	realistic	D	-	-
53	0.334	city-pliers	pltr02wt6	realistic	-	-	-
54	0.328	uva	uamst02wna	realistic	-	A	-
55	0.318	csiro	csiro02np04	realistic	D	A	-
56	0.307	csiro	csiro02np16	realistic	D	A	L
57	0.263	cuny	pirc2wnp1	realistic	D	A	-
58	0.260	uva	uamst02wntm	realistic	-	-	-
59	0.241	csiro	csiro02np02	realistic	-	A	-
60	0.207	umelbourne	mu307	realistic	D	A	-
61	0.150	waterloo	uwmtbw1	realistic	-	-	L
62	0.132	ajou	ajouai0204	realistic	D	-	-
63	0.108	ajou	ajouai0201	realistic	D	-	-
64	0.106	waterloo	uwmtbw4	realistic	-	A	L
65	0.103	waterloo	uwmtbw3	realistic	-	A	L
66	0.076	cuny	pirc2wnp2	realistic	D	A	L
67	0.072	ajou	ajouai0202	realistic	D	-	-
68	0.071	ajou	ajouai0203	realistic	D	-	-
69	0.010	kasetsart	kuhpf0201	realistic	-	A	-
70	0.010	ajou	ajouai0205	realistic	D	-	L