Overview of the TREC-2001 Web Track

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

May 9, 2002

Abstract

TREC-2001 saw the falling into abeyance of the Large Web Task but a strengthening and broadening of activities based on the 1.69 million page WT10g corpus. There were two tasks. The topic relevance task was like traditional TREC ad hoc but used queries taken from real web search logs from which description and narrative fields of a topic description were inferred by the topic developers. There were 50 topics. In the homepage finding task queries corresponded to the name of an entity whose home page (site entry page) was included in WT10g. The challenge in this task was to return all of the homepages at the very top of the ranking.

Cursory analysis suggests that once again, exploitation of link information did not help on the topic relevance task. By contrast, in the homepage finding task, the best performing run which did not make use of either link information or properties of the document's URL achieved only half of the mean reciprocal rank of the best run.

1 Introduction

The TREC-9 Web Track activities centred on two tasks: A Topic Relevance Task and a HomePage Finding Task. Both made use of a 10 gigabyte, 1.69 million document subset of the VLC2, distributed on five CD-ROMs as the WT10g collection. [Bailey et al. 2001].

2 Guidelines

2.1 This Year's Aims

1. To extend the utility of the WT10g Web test collection by obtaining "sufficiently complete" relevance judgements for 50 additional (correctly spelled) ad hoc (topic relevance) topics.
2. To explore a different type of retrieval task (homepage finding) for which it is known that link-based methods can be beneficial.
3. To investigate the benefit (or harm) of correctly implemented link methods on topic relevance.

Participants are welcome to explore specific Web retrieval issues, such as:

1. Can Distributed Information Retrieval techniques be used to improve retrieval effectiveness and/or efficiency?
2. How well can systems accommodate to misspelled queries. Note that the intention is that the standard query set will be correctly spelled so that we maximise the chance of finding all the relevant answers. However, if participants are sufficiently interested, we could issue a set of misspelled variants of the judged queries.

There are obviously many other interesting questions to ask about the Web data.

2.2 Dataset

The data for the TREC-9 Main Web Task is the 10 gigabyte WT10g [CSIRO 2001] collection, distributed by CSIRO. Note that this is entirely Web data. Documents include the information returned by the http daemon (enclosed in DOCHDR tags) as well as the page content. A draft paper [Bailey et al. 2001] describing the WT10g collection is available.

2.3 Web Ad Hoc Task

TREC-2001 ad hoc topics (topics 501-550) were created by NIST. They are available from the main TREC website [National Institute of Standards and Technology 1997]. They take a similar form to previous TREC Ad Hoc topics, but the topic title is a real Web query taken from search engine logs and the other fields are reverse engineered by NIST assessors. The additional fields are intended to define what the searcher wanted (but didn't fully specify) when they typed their query.

Systems are officially compared only on the basis of title-only queries, processed completely automatically. Queries using additional fields have no Web reality! However, despite this, participants were encouraged to submit additional interactive, manual and full topic statement runs to increase the discovery rate of relevant documents in the collection. As part of the automated submission process, participants were required to identify the type of each run.

Official training data (distributed by NIST) consisted of the TREC-9 topics and qrels (topics 451-500). These were directly comparable with the TREC-2001 task.

2.4 Home Page Finding Task

NIST devised a set of 145 homepage finding queries. The process involved finding a homepage within WT10g and then composing a query designed to locate it. This is a known-item search task in which each known item is the entry page to a Website. As an example, the query "Text Retrieval Conference" might be generated for the `http://trec.nist.gov/` homepage. A minimal amount of judging was required to determine if the URLs of documents returned by participants were in fact equivalent to the answer originally chosen. For example, `http://allen.rad.nd.edu:80/` and `http://rad.nd.edu/` both refer to the home page for the Notre Dame Radiation Laboratory.

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

A set of 100 queries and correct answers generated by Nick Craswell using a similar method were made available [CSIRO 2001] for training purposes.

No manual or interactive query modification was permitted in this task. There was a blanket prohibition on tuning, tweaking or altering of systems based on examining the test queries.

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

1. All submissions were due at NIST on or before 2 August 2001.

2. An automated submission process was used which collected a small amount of information about each run.

3. No. of runs submitted/judged.

4. All judging was performed by NIST (not CSIRO) assessors.

5. Judgments in the Web Ad Hoc task (not Homepage Finding) were TERNARY (nonrelevant, relevant, highly relevant) as they were last year.

6. Judgments were made on the basis of the text within the document (only)

7. Judges were not able to follow links.

In the Topic Relevance task, 70400 documents were judged and 3363 were judged either relevant (2573) or highly relevant (790).

In the Homepage Finding task, there were a total of 252 right answers over the 145 queries, an average of 1.74 right answers per query. However, the distribution of number of right answers per query was very skewed. For 132 queries there was only one right answer but for three queries there were more than 10 right answers: query EP33 (Best Internet) - 25, query EP122 (Society for Technical Communications) - 22, and query EP139 (The Leader OnLine) - 17).

Best Internet seems to be (have been) an internet hosting company which controls a whole lot of internet domain names and presents all of them with its own homepage (prior to selling them to customers I presume). The URLs by which this page was accessible included: `www.voici.com`, `www.avantisoft.com`, `www.panint.com`, `www.samoyed.org`, `www.cookiefactory.com`, `www.prost.org`, `www.bayberry.com`, `www.voici.com`, `www.biloxi-ms.com`, `www.globeprint.com`, `www.buoymedia.com`, `www.nm-solutions.com`, `www.growing.com`, `www.caber.com`, `apogee.best.com`, `204.156.149.14`, `www.weblab.com`, `www.anymtnltd.com`, `www.romenet.com`, `www.spottedantelope.com`, `www.straw.com`, `www.jjsblues.com`, `www.jointventure.org`, `204.156.144.1`, `www.mochinet.com`, `www.flick.com`.

By contrast, the multiple results for the Society for Technical Communications, seem to include some spurious answers. The real home page appears to be at `www.stc-va.org/display.html` but lots of the others judged equivalent are subsidiary pages or homepages of individual chapters or regions of STC.

Finally, the multiple answers for the Online Leader, correspond to separate issues of an online publication. Each issue looks like a homepage but each has a specific date, eg. `www.olympus.net/leader/leaderonlineoctober23961023.htm`. The page which you might expect to be a homepage (`www.olympus.net/leader/index.html`) also has a date.

We considering URL depth to be the number of slashes in the URL after eliminating trailing slashes, we computed a histogram of the shallowest right answer for each of the queries. It turns out that 95 of the 145 shallowest answers are actually at the very top level eg. `africa.cis.co.za:81`, `amelia.experiment.db.erau.edu`, `dbcl13.cs.ust.hk01`. Only 11 of the shallowest right answers are at a depth greater than 2.

3 Results

3.1 Topic Relevance Task

Table 1 gives details of the 77 official submissions in the title-only, automatic category of the Topic Relevance task. The best performing run fub01be2 (FUB) did not make use of links, document structure, or URL text. Features listed for that run were: no-stemming, single-word indexing, novel probabilistic term weighting model, automatic query expansion.

The second best run JuruFull (IBM-Haifa) used document structure and referring anchortext. Features listed for that run were: Vector space model, using lexical affinites, Porter stemming, slight stop-word filtering.

The best run from the third-ranked group (Ricoh) used only document content. Features listed for that run were: Probabilistic model, Query expansion, Automatic parameter value estimation

The best run from the fourth ranked group (JustSystem) made use of link information but at this stage it is unclear how. Features listed for that run were: vector space search, reference DB, pseudo-relevance feedback

In summary, it was possible to achieve top performance using document content only. Automatic query expansion was used by most of the top ranked runs. There was no clear advantage to either probabilistic or vector space approaches.

Table 1 presents early precision results for the same official title-only runs.

Table 3 gives details for the 20 other runs, including two manual runs. The best full-topic automatic run performed 27% better than the best title-only run. Interestingly, it made use of URL text as well as page content.

3.2 Home Page Finding Task

Table 4 gives details of all 43 official runs in the Home Page Finding task. Interestingly, the top 23 runs in this table all made use of either URLtext or links (or both). The best run which did not (IBMHOMENR) achieved an MRR score only half as high as that of the top ranked run. It made use of document structure. The highest ranked run which used content only achieved an MRR score only 30% of the best and found a right answer in the top 10 only half as often.

The performance of the top ranked run (tnout10epCAU) is quite impressive. It found a right answer in the top 10 in nearly 90% of cases. The features of this run were listed as follows: Unigram language model URL text priors (based on depth of URL-path) content run merged with seperate anchor-text run. Interestingly, a companion run which did not use anchor text scored almost as well, reflecting the importance of URL depth as a feature on this task - at least for this set of queries on this collection.

Acknowledgements

With assistance from her colleagues at NIST, Ellen Voorhees played a major role in organising the Web track, through topic formulation, assessment, evaluation and analysis. Much of the Main Web data and many of the analyses reported here are the result of her work.

The pivotal contribution of Peter Bailey in engineering the WT10g is gratefully acknowledged.

We are very much indebted to Brewster Kahle of the Internet Archive for making available the spidered data from which the VLC2 and WT10g collections.

Finally, thanks are due to the NIST assessors for topic development and assessments.

Bibliography

BAILEY, P., CRASWELL, N., AND HAWKING, D. 2001. Engineering a multi-purpose test collection for web retrieval experiments. *Information Processing and Management*. In revision. `www.ted.cmis.csiro.au/~dave/cwc.ps.gz`.

CSIRO. 2001. TREC Web Tracks home page. `www.ted.cmis.csiro.au/TRECWeb/`.

NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY. 1997. TREC home page. `trec.nist.gov/`.

Table 1: All official submissions in the title-only, automatic topic relevance task, ranked on average precision.

Runid	Group	Fields	Struct.	URLtext	Links	AveP	ret(100)	ret(1000)
fub01be2	FUB	T	-	-	-	0.2226	17.38	46.1
JuruFull	IBM-Haifa	T	Y	Y	-	0.2105	17.22	45.6
JuruFullQE	IBM-Haifa	T	Y	Y	-	0.2091	16.92	42.6
ricMM	ricoh	T	-	-	-	0.2084	16.84	47.4
ricAP	ricoh	T	-	-	-	0.2077	17.62	49
ricMS	ricoh	T	-	-	-	0.2068	16.84	47.4
JuruPruned01	IBM-Haifa	T	Y	Y	-	0.2066	17.48	43.1
JuruPrune005	IBM-Haifa	T	Y	Y	-	0.2065	17.3	44.1
jscbtawtl4	Justsystem	T	Y	-	Y	0.2060	16.88	46.9
jscbtawtl3	Justsystem	T	Y	-	Y	0.2003	16.9	45.9
Lemur	cmu-lti	T	-	-	-	0.1985	17.58	48
fub01ne2	FUB	T	-	-	-	0.1962	16.42	42.9
jscbtawtl2	Justsystem	T	Y	-	Y	0.1954	16.3	45.2
ok10wt3	microsoft	T	-	Y	-	0.1952	16.86	47.2
hum01tlx	hummingbird	T	Y	-	-	0.1949	16.48	45.8
ricST	ricoh	T	-	-	-	0.1933	16.2	46.3
msrcn1	microsoft-china	T	Y	Y	-	0.1913	15.9	45.1
ok10wt1	microsoft	T	-	Y	-	0.1908	16.78	46.7
fub01idf	FUB	T	-	-	-	0.1900	16.44	42.4
tnout10t2	tno/utwente	T	-	-	-	0.1891	16.66	47.1
iit01tfc	IIT	T	-	-	-	0.1890	16.92	47.6
jscbtawtl1	Justsystem	T	Y	-	Y	0.1890	15.84	44.6
msrcn4	microsoft-china	T	Y	Y	-	0.1880	14.12	43.6
msrcn2	microsoft-china	T	Y	Y	Y	0.1864	14.14	43.6
fub01ne	FUB	T	-	-	-	0.1790	14.98	41.2
hum01tl	hummingbird	T	Y	-	-	0.1784	15.1	43.7
msrcn3	microsoft-china	T	Y	Y	-	0.1779	14.3	39.9
posnir01rpt	postech	T	Y	-	-	0.1771	14.7	42.1
pir1Wt2	cuny	T	-	-	-	0.1742	15.2	45.5
flabxt	Fujitsu	T	-	-	-	0.1719	16.4	43.1
UniNEtdL	Neuchatel	T	-	-	-	0.1715	14.58	43.2
flabxtl	Fujitsu	T	-	-	Y	0.1705	16.06	43
UniNEt7dL	Neuchatel	T	-	-	-	0.1699	14.66	44
fdut10wtc01	Fudan	T	-	-	-	0.1661	15.1	34.7
pir1Wt1	cuny	T	-	-	-	0.1660	14.68	45.3
UniNEtd	Neuchatel	T	-	-	-	0.1659	14.34	43.3
tnout10t1	tno/utwente	T	-	-	-	0.1652	14.62	43.6
hum01t	hummingbird	T	Y	-	-	0.1582	14.42	40.8
apl10wc	apl-jhu	T	-	-	-	0.1567	14.12	42.1
fdut10wtl01	Fudan	T	-	-	Y	0.1544	14.56	34.7
posnir01st	postech	T	Y	-	-	0.1535	13.7	42
posnir01pt	postech	T	Y	-	-	0.1521	13.96	42.2
iit01t	IIT	T	-	-	-	0.1509	13.92	40.1
ARCJ0	ibm-web	T	Y	-	-	0.1497	11.94	31.4
ARCJ5	ibm-web	T	Y	-	Y	0.1439	11.88	31.4
Merxt	IRIT	T	-	-	-	0.1438	13.76	39.9
uwmtaw2	waterloo	T	-	-	-	0.1420	13.88	39.9
uwmtaw1	waterloo	T	-	-	-	0.1416	12.84	39
PDWTAHDR	padova	T	-	-	-	0.1332	12.74	37.8
Ntvenx2	nextrieve	T	Y	-	-	0.1313	11.94	33.3
yeahtb01	Yonsei	T	Y	-	-	0.1287	12.84	26.8
yeaht01	Yonsei	T	Y	-	Y	0.1286	12.82	26.7
Ntvenx1	nextrieve	T	Y	-	-	0.1273	11.76	35.5
PDWTAHWL	padova	T	-	-	Y	0.1209	11.56	37.8
Ntvfnx3	nextrieve	T	Y	-	-	0.1128	11.94	30.1
ajouai0103	ajou	T	-	-	Y	0.1116	10.72	37.1
ajouai0101	ajou	T	-	-	-	0.1114	10.74	37.1
csiro0awa1	CSIRO	T	Y	Y	Y	0.1085	10.58	34.3
uncvsms	uncYang	T	-	-	-	0.1069	12.26	33.4
Ntvfnx4	nextrieve	T	Y	-	-	0.0978	10.08	25.8
uwmtaw0	waterloo	T	-	-	-	0.0951	11.3	27.2
csiro0awa3	CSIRO	T	Y	Y	Y	0.0946	10.76	29.8
icadhoc3	imperial	T	-	-	-	0.0883	9.88	26.9
ictweb10n	chinese_academy	T	-	-	-	0.0860	9.42	28.5
ictweb10nl	chinese_academy	T	-	-	Y	0.0860	9.54	28.5
PDWTAHPR	padova	T	-	-	-	0.0842	10.14	36.7
apl10wa	apl-jhu	T	-	-	-	0.0805	9.72	34
csiro0awa2	CSIRO	T	Y	Y	Y	0.0789	9.48	27.9
apl10wb	apl-jhu	T	-	-	-	0.0671	6.96	12
uncfsls	uncYang	T	-	-	Y	0.0663	11.62	33.3
PDWTAHTL	padova	T	-	-	Y	0.0601	6.82	37.8
icadhoc1	imperial	T	-	-	Y	0.0537	7.96	24.5
ictweb10nf	chinese_academy	T	-	-	-	0.0464	5.68	28.4
ictweb10nfl	chinese_academy	T	-	-	Y	0.0463	5.68	28.4
icadhoc2	imperial	T	-	-	Y	0.0458	8.28	23
irtLnut	uncNewby	T	Y	-	-	0.0221	3.36	16.8
irtLnua	uncNewby	T	Y	-	-	0.0002	0.06	0.6

Table 2: All official submissions in the title-only, automatic topic relevance task, ranked on precision at 10 documents retrieved.

Runid	Group	Fields	Struct.	URLtext	Links	P@5	P@10	P@20
JuruFull	IBM-Haifa	T	Y	Y	-	0.4320	0.3620	0.3130
JuruPrune005	IBM-Haifa	T	Y	Y	-	0.4240	0.3620	0.3090
JuruPruned01	IBM-Haifa	T	Y	Y	-	0.4160	0.3600	0.3070
JuruFullQE	IBM-Haifa	T	Y	Y	-	0.4400	0.3540	0.3160
fub01be2	FUB	T	-	-	-	0.3760	0.3440	0.2860
ricMM	ricoh	T	-	-	-	0.3880	0.3420	0.2950
ricAP	ricoh	T	-	-	-	0.3800	0.3380	0.3000
flabxt	Fujitsu	T	-	-	-	0.3600	0.3360	0.2790
flabxtl	Fujitsu	T	-	-	Y	0.3640	0.3360	0.2770
fub01idf	FUB	T	-	-	-	0.3720	0.3360	0.2880
ok10wt1	microsoft	T	-	Y	-	0.3600	0.3360	0.2950
ok10wt3	microsoft	T	-	Y	-	0.3840	0.3360	0.3040
ricMS	ricoh	T	-	-	-	0.3800	0.3360	0.2950
tnout10t2	tno/utwente	T	-	-	-	0.3640	0.3340	0.2720
hum01tlx	hummingbird	T	Y	-	-	0.3840	0.3320	0.2860
fub01ne2	FUB	T	-	-	-	0.3760	0.3280	0.2760
ricST	ricoh	T	-	-	-	0.3640	0.3260	0.2720
fub01ne	FUB	T	-	-	-	0.3800	0.3240	0.2720
yeaht01	Yonsei	T	Y	-	Y	0.3880	0.3240	0.2520
yeahtb01	Yonsei	T	Y	-	-	0.3840	0.3240	0.2520
hum01tl	hummingbird	T	Y	-	-	0.3680	0.3220	0.2660
Lemur	cmu-lti	T	-	-	-	0.3720	0.3200	0.2920
msrcn2	microsoft-china	T	Y	Y	Y	0.3600	0.3160	0.2620
msrcn4	microsoft-china	T	Y	Y	-	0.3560	0.3160	0.2620
jscbtawtl4	Justsystem	T	Y	-	Y	0.3440	0.3140	0.2830
jscbtawtl3	Justsystem	T	Y	-	Y	0.3360	0.3120	0.2750
fdut10wtl01	Fudan	T	-	-	Y	0.3920	0.3100	0.2410
hum01t	hummingbird	T	Y	-	-	0.3960	0.3080	0.2600
msrcn1	microsoft-china	T	Y	Y	-	0.3520	0.3080	0.2740
msrcn3	microsoft-china	T	Y	Y	-	0.3320	0.3060	0.2660
fdut10wtc01	Fudan	T	-	-	-	0.3840	0.3020	0.2500
iit01t	IIT	T	-	-	-	0.3320	0.3000	0.2470
jscbtawtl2	Justsystem	T	Y	-	Y	0.3360	0.3000	0.2760
jscbtawtl1	Justsystem	T	Y	-	Y	0.2960	0.2980	0.2530
Merxt	IRIT	T	-	-	-	0.3160	0.2880	0.2480
posnir01rpt	postech	T	Y	-	-	0.3640	0.2880	0.2480
iit01tfc	IIT	T	-	-	-	0.3080	0.2840	0.2590
tnout10t1	tno/utwente	T	-	-	-	0.3360	0.2780	0.2420
UniNEtd	Neuchatel	T	-	-	-	0.3120	0.2720	0.2400
PDWTAHDR	padova	T	-	-	-	0.3120	0.2680	0.2220
UniNEtdL	Neuchatel	T	-	-	-	0.3080	0.2660	0.2430
ARCJ0	ibm-web	T	Y	-	-	0.3080	0.2640	0.2210
ARCJ5	ibm-web	T	Y	-	Y	0.2840	0.2640	0.2130
UniNEt7dL	Neuchatel	T	-	-	-	0.3200	0.2640	0.2470
uwmtaw1	waterloo	T	-	-	-	0.2760	0.2580	0.2110
Ntvenx1	nextrieve	T	Y	-	-	0.2720	0.2540	0.2080
PDWTAHWL	padova	T	-	-	Y	0.3080	0.2540	0.2070
posnir01st	postech	T	Y	-	-	0.2960	0.2480	0.2220
Ntvenx2	nextrieve	T	Y	-	-	0.2840	0.2460	0.2080
posnir01pt	postech	T	Y	-	-	0.3040	0.2460	0.2110
csiro0awa1	CSIRO	T	Y	Y	Y	0.2760	0.2440	0.2190
uncvsms	uncYang	T	-	-	-	0.2480	0.2400	0.1920
apl10wc	apl-jhu	T	-	-	-	0.2520	0.2380	0.2200
csiro0awa3	CSIRO	T	Y	Y	Y	0.2200	0.2320	0.1940
pir1Wt1	cuny	T	-	-	-	0.2360	0.2220	0.2070
uwmtaw2	waterloo	T	-	-	-	0.2320	0.2220	0.2130
Ntvfnx4	nextrieve	T	Y	-	-	0.2480	0.2180	0.1770
pir1Wt2	cuny	T	-	-	-	0.2320	0.2160	0.2110
Ntvfnx3	nextrieve	T	Y	-	-	0.2200	0.2120	0.1920
csiro0awa2	CSIRO	T	Y	Y	Y	0.2200	0.2000	0.1770
ajouai0101	ajou	T	-	-	-	0.2040	0.1980	0.1560
ajouai0103	ajou	T	-	-	Y	0.2040	0.1960	0.1570
apl10wb	apl-jhu	T	-	-	-	0.2400	0.1900	0.1580
uwmtaw0	waterloo	T	-	-	-	0.2000	0.1860	0.1700
icadhoc3	imperial	T	-	-	-	0.2080	0.1780	0.1730
ictweb10n	chinese_academy	T	-	-	-	0.1640	0.1600	0.1450
PDWTAHPR	padova	T	-	-	-	0.1680	0.1600	0.1460
ictweb10nl	chinese_academy	T	-	-	Y	0.1600	0.1580	0.1410
apl10wa	apl-jhu	T	-	-	-	0.1600	0.1460	0.1380
icadhoc2	imperial	T	-	-	Y	0.1400	0.1400	0.1370
icadhoc1	imperial	T	-	-	Y	0.1400	0.1260	0.1220
PDWTAHTL	padova	T	-	-	Y	0.1400	0.1140	0.0900
uncfsls	uncYang	T	-	-	Y	0.0600	0.0760	0.1360
ictweb10nf	chinese_academy	T	-	-	-	0.0680	0.0620	0.0620
ictweb10nfl	chinese_academy	T	-	-	Y	0.0680	0.0620	0.0610
irtLnut	uncNewby	T	Y	-	-	0.0400	0.0440	0.0470
irtLnua	uncNewby	T	Y	-	-	0.0000	0.0000	0.0010

Table 3: All other (manual and long automatic) official submissions in the topic relevance task. Manual runs are marked with an asterisk.

Runid	Group	Fields	Struct.	URLtext	Links	AveP	ret(100)	ret(1000)
iit01m*	IIT		-	-	-	0.3324	20.8	43.2
ok10wtnd1	microsoft	TDN	-	Y	-	0.2831	22.36	53.8
csiro0mwa1*	CSIRO		Y	Y	Y	0.2817	19.68	42
ok10wtnd0	microsoft	TDN	-	Y	-	0.2512	20.42	51.7
flabxtd	Fujitsu	TD	-	-	-	0.2332	19.88	49
UniNEn7d	Neuchatel	TDN	-	-	-	0.2242	17.52	48.8
hum01tdlx	hummingbird	TD	Y	-	-	0.2201	18.62	49.4
kuadhoc2001	kasetsart	TDN	-	-	-	0.2088	17.7	44.9
apl10wd	apl-jhu	TDN	-	-	-	0.2035	19.56	50.5
posnir01ptd	postech	TD	Y	-	-	0.1877	17.62	44.5
flabxtdn	Fujitsu	TDN	-	-	-	0.1843	17.32	43.4
iit01tde	IIT	TD	-	-	-	0.1791	16.9	45
Merxtd	IRIT	TD	-	-	-	0.1729	15.58	42.4
pir1Wa	cuny	TDN	-	-	-	0.1715	14.88	45.7
fdut10wac01	Fudan	TDN	-	-	-	0.1661	15.1	34.7
uncvsmm	uncYang	TD	-	-	-	0.1269	14.4	35.9
fdut10wal01	Fudan	TDN	-	-	Y	0.1248	12.72	34.7
yeahdb01	Yonsei	TD	Y	-	-	0.1094	11.52	23.5
yeahtd01	Yonsei	TD	Y	-	Y	0.1092	11.48	23.5
uncfslm	uncYang	TD	-	-	Y	0.0781	13.46	35.8

Table 4: All official submissions in the homepage finding task. MRR is the mean reciprocal rank of the first correct answer. %top10 is the proportion of queries for which a right answer was found in the top 10 results. %fail is the proportion of queries in which no right answer was found in the top 100 results.

Runid	Baseline	Group	Struct.	URLtext	Links	MRR	%top10	%fail
tnout10epCAU	tnout10epCU	tno/utwente	-	Y	Y	0.774	88.3	4.8
tnout10epCU		tno/utwente	-	Y	-	0.772	87.6	4.8
jscbtawep2		Justsystem	Y	Y	Y	0.769	83.4	9.0
jscbtawep1		Justsystem	Y	Y	Y	0.754	83.4	9.0
jscbtawep4		Justsystem	Y	Y	Y	0.752	83.4	8.3
jscbtawep3		Justsystem	Y	Y	Y	0.746	83.4	9.0
yehp01	yehpb01	Yonsei	Y	Y	Y	0.669	76.6	22.1
yehpb01		Yonsei	Y	Y	-	0.659	75.9	22.8
UniNEep1		Neuchatel	-	Y	-	0.637	69.0	8.3
UniNEep2		Neuchatel	-	Y	-	0.637	69.0	7.6
IBMHOMER	IBMHOMENR	ibm-web	Y	-	Y	0.611	77.9	10.3
flabxeall		Fujitsu	-	-	Y	0.599	80.7	9.7
csiro0awh2		CSIRO	-	-	Y	0.593	71.7	21.4
iit01stb	iit01st	IIT	Y	Y	Y	0.578	66.9	24.8
iit01st		IIT	Y	Y	-	0.559	62.8	29.7
UniNEep3		Neuchatel	-	Y	-	0.530	68.3	6.9
VTEP	VTBASE	VT	-	Y	Y	0.506	68.3	15.9
msrcnp2	msrcnp1	microsoft-china	Y	Y	Y	0.505	69.0	15.2
csiro0awh1	csiro0awh3	CSIRO	Y	Y	Y	0.498	72.4	11.0
UniNEep4		Neuchatel	-	Y	-	0.477	68.3	11.0
msrcnp1		microsoft-china	Y	Y	-	0.424	65.5	13.1
flabxe75a		Fujitsu	Y	Y	Y	0.399	55.9	37.9
ok10wahd1	ok10whd1	microsoft	-	Y	Y	0.387	64.1	13.1
IBMHOMENR		ibm-web	Y	-	-	0.382	62.1	11.7
flabxemerge		Fujitsu	Y	Y	Y	0.365	51.0	33.8
flabxet256		Fujitsu	Y	-	Y	0.363	50.3	33.8
ok10wahd0	ok10whd0	microsoft	-	Y	Y	0.362	62.1	13.1
ok10whd1		microsoft	-	Y	-	0.340	60.7	15.9
tnout10epC		tno/utwente	-	-	-	0.338	58.6	18.6
tnout10epA		tno/utwente	-	-	Y	0.331	48.3	35.9
ok10whd0		microsoft	-	Y	-	0.312	58.6	15.2
apl10ha		apl-jhu	-	-	-	0.238	44.8	22.1
ichp2		imperial	-	-	-	0.237	44.8	29.7
apl10hb		apl-jhu	-	-	-	0.220	42.8	21.4
ichp1	ichp2	imperial	-	-	Y	0.208	33.8	37.2
kuhpf2001		kasetsart	-	-	-	0.191	36.6	42.1
PDWTEPDR		padova	-	-	-	0.189	33.8	42.8
PDWTEPWL	PDWTEPDR	padova	-	-	Y	0.178	30.3	42.8
VTBASE		VT	-	-	-	0.126	24.1	45.5
ajouai0102		ajou	-	-	-	0.101	23.4	49.7
ajouai0104		ajou	-	-	Y	0.100	23.4	49.7
PDWTEPTL	PDWTEPDR	padova	-	-	Y	0.099	20.0	42.8
PDWTEPPR		padova	-	-	-	0.054	13.1	44.8