Overview of the TREC 2013 Federated Web Search Track

Thomas Demeester[1], Dolf Trieschnigg[2], Dong Nguyen[2], Djoerd Hiemstra[2]
[1] Ghent University - iMinds, Belgium
[2] University of Twente, The Netherlands
tdmeeste@intec.ugent.be, {d.trieschnigg, d.nguyen, d.hiemstra}@utwente.nl

ABSTRACT

The TREC Federated Web Search track is intended to promote research related to federated search in a realistic web setting, and hereto provides a large data collection gathered from a series of online search engines. This overview paper discusses the results of the first edition of the track, FedWeb 2013. The focus was on basic challenges in federated search: (1) resource selection, and (2) results merging. After an overview of the provided data collection and the relevance judgments for the test topics, the participants' individual approaches and results on both tasks are discussed. Promising research directions and an outlook on the 2014 edition of the track are provided as well.

1. INTRODUCTION

Building large-scale search engines increasingly depends on combining search results from multiple sources. A web search engine might combine results from numerous verticals, such as: videos, books, images, scientific papers, shopping, blogs, news, recipes, music, maps, advertisements, Q& A, jobs, social networks, etc. Typically, the search results provided by each source differ significantly in the provided snippets, the provided additional (structured) information, and the ranking approach used. For online shopping, for instance, the results are highly structured, and price, bids, ratings and click-through rate are important ranking criteria, whereas for scientific paper search the number of citations is an important ranking criterion. Federated search also enables the inclusion of results from otherwise hidden web collections that are not easily crawlable.

The TREC Federated Web Search (FedWeb) track 2013 provides a test collection that stimulates research in many areas related to federated search, including aggregated search, distributed search, peer-to-peer search and meta-search engines [19]. The collection relieves researchers from the burden of collecting or creating proprietary datasets [3], or creating artificial federated search test collections by dividing existing TREC collections by topic or source [14]. The TREC FedWeb 2013 collection is different from such artificially created test collections in that it provides the actual results of 157 real web search engines, each providing their own retrieval method and heterogeneous content types including images, pdf-text, video, etc. [2]. This paper describes the first edition of the TREC FedWeb track. A total of 11 groups (see Table 1) participated in the two classic distributed search tasks [9]:

Task 1: Resource Selection
The goal of resource selection is to select the right resources from a large number of independent search engines given a query. Participants had to rank the 157 search engines for each test topic without access to the corresponding search results. The FedWeb 2013 collection contains search result pages for many other queries, as well as the HTML of the corresponding web pages. These data could be used by the participants to build resource descriptions. Some of the participants also used external sources such as Wikipedia, ODP, or WordNet.

Task 2: Results Merging
The goal of results merging is to combine the results of several search engines into a single ranked list. After the deadline for Task 1 passed, the participants were given the search result pages of 157 search engines for the test topics. The result pages include titles, snippet summaries, hyperlinks, and possibly thumbnail images, all of which were used by participants for reranking and merging. In later editions of the track, these data will also be used to build aggregated search result pages.

The official track guidelines can be found online[1].

Apart from studying resource selection and results merging in a web context, there are also new research challenges that readily appear, and for which the FedWeb 2013 collection could be used. Some examples are: How does the snippet quality influence results merging strategies? How well can the relevance of results be estimated based on snippets only? Can the size or the importance of search engines be reliably estimated from the provided search samples? Are people able to detect duplicate results, i.e., the same result provided by multiple search engines?

This overview paper is organized as follows: Section 2 describes the FedWeb collection; Section 3 describes the process of gathering relevance judgements for the track; Sections 4 and 5 describe the results for the resource selection task and results merging task, respectively; Section 6 gives a summary of this year's track and provides an outlook on next year's track.

[1] http://snipdex.org/fedweb

Group ID	Institute	RS runs	RM runs
CWI	Centrum Wiskunde & Informatica	3	3
ICTNET	Chinese Avademy of Sciences		3
IIIT_Hyderabad	International Institute of Information Technology	1	
NOVASEARCH	Universidade Nova de Lisboa		3
isi_pal	Indian Statistical Institute	2	1
scunce	East China Normal University	1	
StanfordEIG	Stanford University	1	
udel	University of Delaware	3	3
UiS	University of Stavanger	3	
UPD	University of Padova	2	2
ut	University of Twente	2	

Table 1: Participants and number of runs for Resource Selection (RS) and Results Merging (RM).

		Total	Per engine
Samples (2000 queries)	Snippets	1,973,591	12,570.6
	Pages	1,894,463	12,066.6
	Size (GB)	177.8	1.13
Topics (200 queries)	Snippets	143,298	912.7
	Pages	136,103	866.9
	Size (GB)	16.7	0.11

Table 2: FedWeb 2013 collection statistics

Category	Count	Category	Count
Academic	18	Local	1
Audio	6	News	15
Blogs	4	Photo/Pictures	13
Books	5	Q&A	7
Encyclopedia	5	Recipes	5
Entertainment	4	Shopping	9
Games	6	Social	3
General	6	Software	3
Health	12	Sports	9
Jobs	5	Tech	8
Jokes	2	Travel	2
Kids	10	Video	14

Table 3: FedWeb 2013 search engine categories (an engine can be in multiple categories)

2. FEDWEB 2013 COLLECTION

The FedWeb 2013 Data Collection consists of search results from 157 web search engines in 24 categories ranging from news, academic articles and images to jokes and lyrics. Overview statistics of the collection are listed in Table 2. The categories are listed in Table 3, and the search engines are listed in Appendix A. To prevent a bias towards large general web search engines, we merged the results from a number of large web search engines into the 'BigWeb' (engine e200) search engine. A query for this engine was sent randomly to one of the large web search engines. In comparison to the 2012 collection (available for training) [17], the 2013 collection covers more search engines and a larger variety of categories and has more samples. The collection contains both the search result snippets and the pages the search results link to.

2.1 Extracting snippets

The search result snippets were scraped from the HTML search result pages using XPaths. This allowed a single approach to be used for all engines rather than to program a wrapper for each search engine API. The SearchResultFinder plugin [21, 20] was used to quickly identify reusable XPaths to extract the snippets from search result pages. Additional (relative) XPaths were determined manually to extract the link, title, description and thumbnail from each snippet. Table 4 shows an example of the required information to sample search results from a single search engine. Up to 10 snippets from the first search result page were extracted for each engine.

2.2 Sampling

2000 sample queries were issued to each of the 157 search engines. The first set of a 1000 queries was the same across all search engines and were single words sampled from the vocabulary of the ClueWeb09-A collection. The second set of a 1000 queries was engine-dependent and consisted of single words sampled from the retrieved snippet vocabulary of that engine. The pages and thumbnails that were linked to from the snippets were downloaded and included in the collection.

2.3 Topics

The organizers created 200 topic descriptions and queries, targeted at specific categories in the collection. Similar to the sampling, for each of the topics the top 10 search result snippets and pages from each search engine were crawled. To facilitate the judgements of pages, screenshots were taken using Selenium[2] (with a maximum height of 3000 pixels) of the top of each retrieved page.

2.4 Duplicate page detection

The search engines in the collection have overlapping indexes, which might result in duplicate pages in the merged search results. To prevent rewarding merged search results containing duplicate (relevant) content, we semi-automatically determined duplicate content. First, a set of candidate duplicates was determined automatically. Then, pairs of likely duplicates were checked manually to determine their state.

Pairs of pages were considered duplicate when:

[2] `http://seleniumhq.org`

Search engine	University of Twente (e014)
Search URL	`http://doc.utwente.nl/cgi/search/simple?q={q}`
Item XPath	`//tr[@class='ep_search_result']`
Title XPath	`.//em`
Description XPath	`.`
Link XPath	`.//a/@href`
Thumbnail XPath	`.//img[@class='ep_preview_image']/@src`

Table 4: Example XPaths for scraping snippets from result pages

1. Their normalized URLs are the same. The URL is normalized by lowercasing it, removing the `www.` prefix of a URL, replacing `https` by `http`, removing trailing slashes and paths ending with `index.html` and `index.php`.

2. The pages are not empty and their MD5 hashes are the same.

3. Both URLs do not appear on a manually compiled exclusion list which are known to contain false positives (e.g. from phdcomics.com), the pages contain at least 100 words, have a similar length ($< 2\%$ difference) and have the same Simhash [11].

The pairs of pages in the third category were manually checked. False positives included URLs that simply showed a "not available anymore" page and pages asking to accept cookies to view the page. 12,903 pages were flagged as duplicate, resulting in 4,601 page types.

3. RELEVANCE ASSESSMENTS

This section describes the collection of the test topics and the relevance judgments, and gives an idea of how the different resource categories contribute to the total fraction of relevant results.

To collect test topics, we first created a pool of new queries and queries from previous TREC tracks (all queries from the Web Track 2009 and 2010, and selected queries from the Million Query Track 2009). The 271 new queries are real life queries, recorded by a number of people with diverse backgrounds, who provided both the queries and the corresponding information need descriptions. We explicitly asked them to also include queries targeting other than only general web search engines. For all 506 queries in this pool, we estimated which resource categories (see Table 3) each of those queries was most likely to target, and made a first selection of 200 (mostly new) queries, thereby ensuring that all resource categories were well represented. The annotation was then organized in two distinct steps. First, we judged all top-3 snippets from each resource for each of these 200 queries (in total almost 50,000 snippets), given that judging snippets goes much faster than judging pages. From those 200 queries, we selected 50 queries for which we collected the complete page judgments (i.e., for the top 10 results). These 50 queries were selected based on the relevance distribution of the judged snippets, avoiding queries with too few or too many relevant results. We also favored queries which had a substantial number of relevant results among other than only the general web search engines. For those 50 queries, the judges were asked to write down a *narrative* which described the information need, its context and the expected results. This narrative was used in both the snippet and page judgments. We collected over 32,000 page judgments for the 50 selected queries, not including overlapping judgments. An example of a query, with description and narrative, is given below.

```
<topic id='7145'>
  <query>why do cats purr</query>
  <description>
    You want to know why cats purr and what it means.
  </description>
  <narrative>
    You have cats and want to know what they want to
    communicate while purring. Any information on the
    meaning of purring is interesting, including videos.
    However, biological information on how the purring
    is done, is not relevant.
  </narrative>
</topic>
```

The graded relevance levels used in the judgements are also used in the Web Track[3]: Non (not relevant), Rel (minimal relevance), HRel (highly relevant), Key (top relevance), and Nav (navigational).

There are a number of differences with respect to the 2012 test collection [17]. First of all, we judged *all* pages (in the top 10 result lists), whereas for the 2012 test topics we left out those with non-relevant snippets. Also, besides the information need descriptions, we introduced a narrative for each query, facilitating the assessor's consistent choice of relevance for results from different resource categories. The main difference is, however, the choice of test queries designed to avoid the strong bias towards general web search engines, mentioned in [12]. As a reference, we added the 50 selected queries in Appendix B. As an illustration, Fig. 1 gives an overview of the relevance distribution over the different resource categories, in a boxplot that presents per category the fraction of results with relevance level Rel or higher for each test topic. For the most important resource categories (in terms of number or size of resources, i.e: General, Video, Blogs, Audio, ...), many topics provide a significant amount of relevant results. However, we also tried to select at least a few topics targeting smaller resource categories (e.g., Recipes, Travel, Jokes). In the end, only two categories (Games and Local) did not provide a notable number of relevant results for any of the test topics, despite queries that were intended to target those categories, like query 7415 ('most anticipated games of 2013') for games, or 7009 ('best place to eat pho in new york') for local, see appendix B.

[3] `http://research.microsoft.com/en-us/projects/trec-web-2013/`

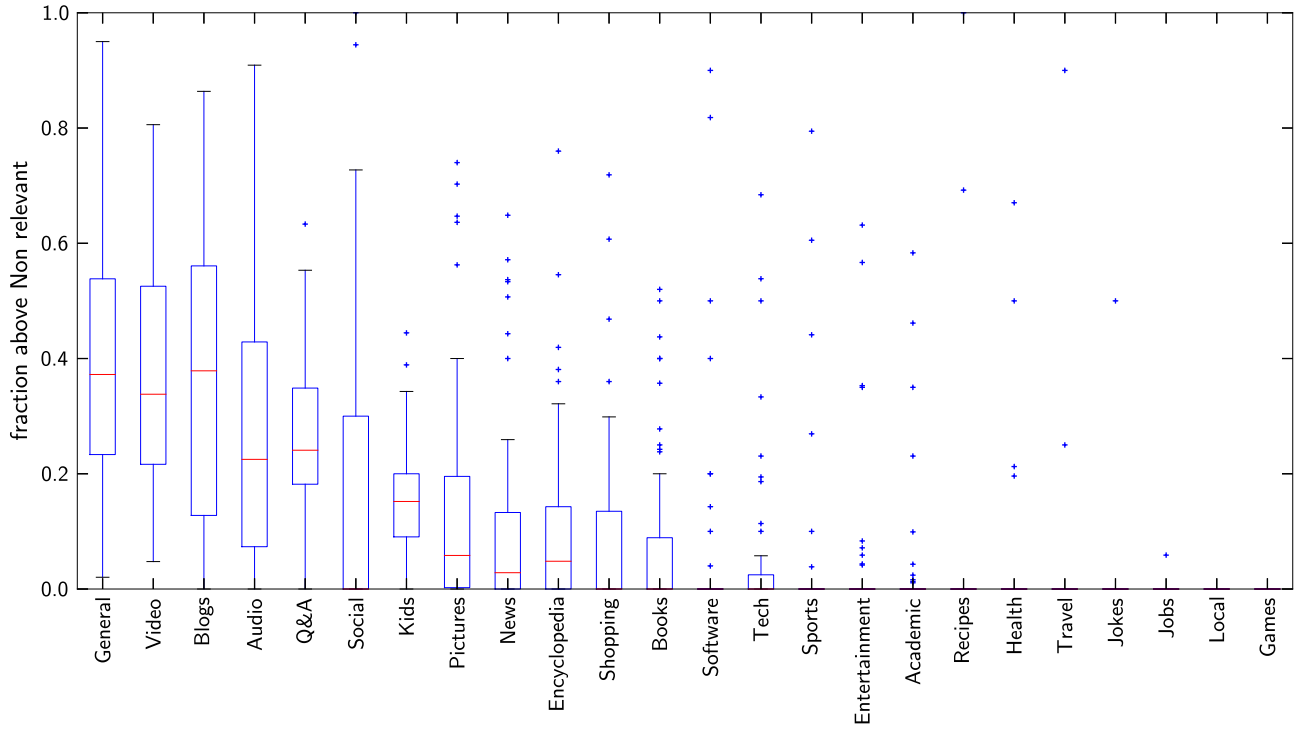

Figure 1: Relevance distributions over resource categories.

4. RESOURCE SELECTION

4.1 Evaluation

The evaluation results for the resource selection task are shown in Table 5, displaying for a number of metrics the average per run over all topics. The primary evaluation metric for the resource selection task is the normalized discounted cumulative gain nDCG@20, where we use the nDCG variant introduced by Burges et al. [8]. The gain g_j at rank j is calculated as $g_j = 2^{r(j)} - 1$, with $r(j)$ the relevance level of the result at rank j. The relevance of a search engine for a given query is determined by calculating the graded precision [15] on the top 10 results. This takes the graded relevance levels of the documents in the top 10 into account, but not the ranking. The following weights are given to the relevance levels of documents: $w_{\text{Non}} = 0, w_{\text{Rel}} = 0.25, w_{\text{HRel}} = 0.5, w_{\text{Key}} = r_{\text{Nav}} = 1$. The graded relevance values are then converted to discrete relevance levels r through multiplication by 100 and taking the nearest integer value. We also reported nP@1 and nP@5, the normalized graded precision for the highest ranked resource, respectively, the top 5 resources, averaged over all topics. We define the normalized graded precision nP@k for each topic as the graded precision on all results for that topic from the top k resources (using the graded relevance weights defined above, and disregarding the ranking of results and resources), normalized by the graded precision of the top k resources for the best possible ranking for that topic. For example, nP@1 denotes the graded precision of the highest ranked resource, divided by the highest graded precision by any of the resources for that topic.

4.2 Participant Approaches

This section shortly describes the experiments by the FedWeb participants for the resource selection task.

University of Delaware (udel)

Resources were ranked based on the average document scores (`udelFAVE`), the rank of the highest ranking document (`udelRSMIN`) and by using rankings of documents to find resource scores with a cut-off (`udelODRA`). Weights and cut-off values were determined from experiments on the FedWeb 2012 dataset.

University of Padova (UPD)

The University of Padova, explored the effectiveness of the TWF-IRF weighting scheme in a Federated Web Search setting [7]. The `UPDFW13sh` run was obtained by combining the query keywords using OR. The `UPDFW13mu` run was created by appending three ranked lists of search engines: First, the engines were returned matching an AND query, then the engines matching an OR query (and not included in the first list) and finally the remaining engines (ordered by id).

University of Twente (ut)

The University of Twente used the recently proposed shard selection method called 'Taily' that is based on statistics of all shards [1]. As these were not available, document samples were used instead. In comparison with their original publication, the FedWeb submission assumed that all resources are of the same size. They experimented with a baseline run (`utTailyM400`), and a variation using a Gaussian distribution instead of a Gamma distribution (`utTailyNormM400`).

Task 1: Resource Selection Group ID	Run ID	**nDCG@20**	nP@1	nP@5	resources used
UPD	UPDFW13mu	0.299	0.16	0.21	documents
	UPDFW13sh	0.247	0.12	0.21	documents
UiS	UiSP	0.276	0.18	0.27	documents
	UiSSP	0.274	0.19	0.29	snippets + documents
	UiSS	0.165	0.16	0.21	snippets
udel	udelFAVE	0.244	0.20	0.22	documents
	udelODRA	0.159	0.21	0.18	documents
	udelRSMIN	0.053	0.06	0.07	documents
ut	utTailyM400	0.216	0.17	0.23	documents
	utTailyNormM400	0.214	0.20	0.23	documents
CWI	cwi13SniTI	0.123	0.10	0.19	snippets
	cwi13ODPTI	0.096	0.14	0.16	snippets + ODP
	cwi13ODPJac	0.050	0.06	0.09	ODP
III_Hyderabad	iiitnaive01	0.107	0.13	0.17	snippets, Wikipedia, WordNet
scunce	ECNUBM25	0.105	0.07	0.10	snippets, Google search
isi_pal	incgqdv2	0.037	0.11	0.06	GoogleQuery
	incgqd	0.025	0.09	0.03	GoogleQuery
StanfordEIG	StanfordEIG10	0.018	0.07	0.02	documents
organizers (baselines)	RS_clueweb	0.298	0.00	0.32	snippets
	RS_querypools	0.185	0.07	0.10	

Table 5: Results for the Resource Selection task.

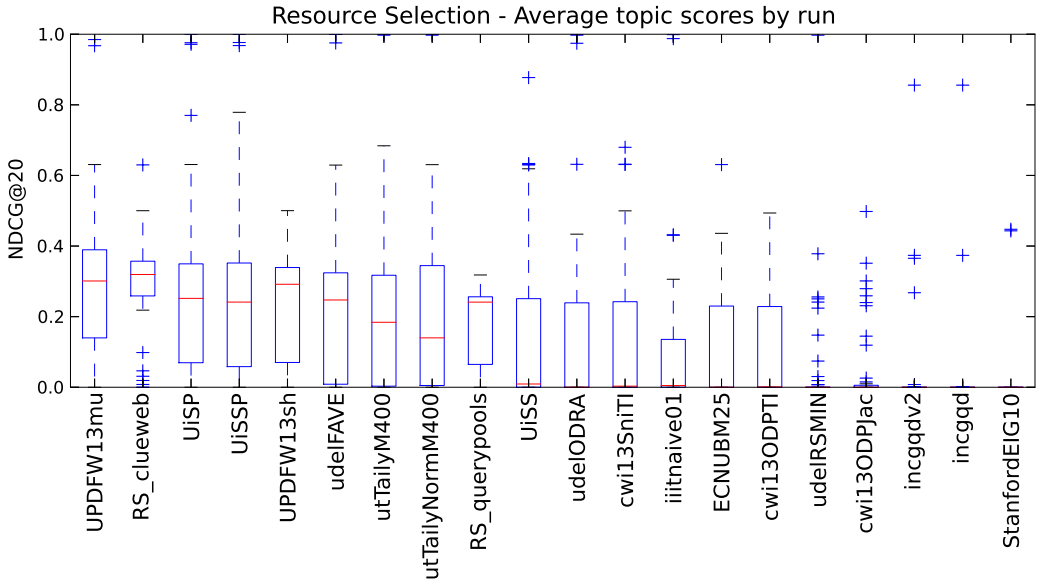

Figure 2: Comparison of runs on the Resource Selection task.

Centrum Wiskunde & Informatica (CWI)

CWI [5] explored the use of ODP category information for resource selection, by ranking the resources based on the Jaccard similarity between the ODP categories of the query and each resource (`cwi13ODPJac`). They also experimented with an approach using only snippets. An index was created of large documents, each created by concatenating all snippets from a resource. The resources were then ranked based on TF-IDF similarity (`cwi13SniTI`). The `cwi13ODPTI` combined the rankings of the two approached using a Borda voting mechanism.

University of Stavanger (UiS)

The University of Stanvanger explored two different approaches [4]. The `UiSP` run ranked individual documents in a central index of all sampled documents, based on their full page content using a language modeling approach. The relevance estimates were then aggregated on a resource level. The `UiSPP` run is a linear combination of the `UiSP` run with a model that estimated the relevance of collections based on a language modeling approach, by representing each resource as a single, large document created from the sampled snippets. `UiSS` used the same approach as `UiSPP`, but now using only snippets. Resource priors were calculated based on the total number of sampled documents.

International Institute of Information Technology (IIIT_Hyderabad)

IIIT Hyderabad explored the use of Wordnet synonyms and Wikipedia categories for query expansion (`iiitnaive01`).

East China Normal University (scunce)

They performed query expansion using Google search and ranked the resources based on BM25 (`ECNUBM25`).

Indian Statistical Institute (isi_pal)

The Indian Statistical Institute did not use the provided document and snippet samples (runs `incgqd` and `incgqdv2`) [18]. Instead, they used the Google Search API to issue the test queries to each resource. Each resource was ranked using the top 8 retrieved results.

Stanford University (StanfordEIG)

The `StanfordEIG10` run was executed over a Cassandra database containing meta information about the search engines. The overall dataset was partitioned into Solr indexes, vectors were then calculated on a TF-IDF basis which was loaded into a dictionary map. Thresholds for term frequency were established at $\geq$ 10,$\geq$ 50 and $\geq$ 100 respectively. Queries were tokenized before being executed over keys and fields in the Cassandra Keyspace. Unfortunately the scoring metric was not stable and only the top result for each query was presented.

4.2.1 Organizers' baseline

As a simple baseline, we used a query-independent method by ranking resources based on their estimated size. The first size estimation method (`RS_clueweb`) scaled the document frequencies in the sampled data based on a reference corpus, for which we used the ClueWeb09 collection[4]. The second method used query pools, similar to [6], and resulted in moderate baseline results (`RS_querypools`).

[4] `http://lemurproject.org/clueweb09/`

4.3 Analysis

Table 5 lists the particpants' results on the Resource Selection task. The NDCG@20 scores range from 0.025 to 0.295 and are strongly correlated (Pearson's r = 0.9) with the nP@5. Fig. 2 visualizes the topic scores per run: a boxplot shows the first and third quartiles and the median (red line) NDCG@20 values.

The Clueweb09 baseline `RS_clueweb` performs surprisingly well. Having a good size estimate turns out to give a solid baseline. Notable is the nP@1 of 0, caused by a flaw in estimating the size of a single search engine (which for every query returns the same set of results). Despite this flaw, the run achieves the highest nP@5. Its boxplot in Fig. 2 shows stable results, with relatively few positive outliers compared to the best peforming run. The other baseline (`RS_querypools`) performs much worse, but similar to `RS_clueweb` it gives relatively stable results with a high median.

The best performing runs (`UPDFW13mu`, `UiSP` and `udelFAVE`) rely on indices based on single documents (rather than snippets) and combine evidence from standard retrieval approaches (variations on TF.IDF and language modeling). The best performing runs do not use external resources such as Wordnet and Wikipedia. A notable exception is the `RS_clueweb` baseline, which uses the collections' snippets in combination with the ClueWeb '09 collection to make size estimates.

5. RESULTS MERGING

5.1 Evaluation

The evaluation results for the results merging task are shown in Table 6, displaying for a number of metrics the average per run over all topics.

The primary evaluation metric for the results merging task is again the normalized discounted cumulative gain nDCG@20. We have chosen the relevance levels used to calculate the gain as $r_{\text{Non}} = 0, r_{\text{Rel}} = 1, r_{\text{HRel}} = 2, r_{\text{Key}} = r_{\text{Nav}} = 3$. Note that when going through the ranked results, duplicate documents (based on URL and content, see Section 2.4) of a result already seen higher in the list, are considered non-relevant (i.e., are assigned relevance level Non), when calculating this measure on the merged results.

We also reported nDCG@100, P@10, and ERR@20, using the same penalty for duplicates. P@10 is the binary precision at 10, whereby all levels from Rel and above are considered relevant, and hence represents the ability of filtering out non-relevant results. For the expected reciprocal rank ERR@20 (see Chapelle et al. [10]), we used the same relevance levels used in the TREC Web Track (i.e., 0-4 ranging from Non to Nav). In order to show that detecting duplicates is an important issue for efficient results merging in the Web setting, we also reported the nDCG@20 and ERR@20 without duplicate penalty, indicated with (*) in Table 6.

5.2 Participant Approaches

Universidade Nova de Lisboa (NOVASEARCH)

NovaSearch experimented with three different late-fusion approaches [16]. Duplicate documents were assumed to have

Task 2: Results Merging Group ID	Run ID	**nDCG@20**	nDCG@100	P@10	ERR@20	nDCG@20(*)	ERR@20(*)
NOVASEARCH	nsRRF	0.257	0.255	0.370	0.254	0.439	0.428
	nsISR	0.165	0.199	0.310	0.166	0.287	0.285
	nsCondor	0.135	0.199	0.278	0.133	0.174	0.171
ICTNET	ICTNETRun2	0.223	0.341	0.414	0.213	0.290	0.274
	ICTNETRun3	0.223	0.322	0.414	0.213	0.290	0.273
	ICTNETRun1	0.216	0.329	0.396	0.206	0.286	0.270
udel	udelRMIndri	0.200	0.369	0.332	0.190	0.366	0.347
	udelSnLnSc	0.161	0.257	0.318	0.159	0.255	0.251
	udelPgLnSc	0.154	0.234	0.318	0.151	0.252	0.244
CWI	CWI13IndriQL	0.162	0.332	0.322	0.154	0.247	0.236
	CWI13iaTODPJ	0.151	0.281	0.284	0.147	0.205	0.200
	CWI13bstTODPJ	0.147	0.240	0.250	0.144	0.230	0.225
UPD	UPDFW13rrmu	0.135	0.170	0.254	0.133	0.231	0.228
	UPDFW13rrsh	0.129	0.171	0.254	0.127	0.222	0.219
isi_pal	merv1	0.081	0.108	0.150	0.081	0.132	0.131
organizers (baselines)	RM_clueweb	0.142	0.260	0.262	0.140	0.167	0.164
	RM_querypools	0.064	0.196	0.186	0.063	0.060	0.058

Table 6: Results for the Results Merging task.

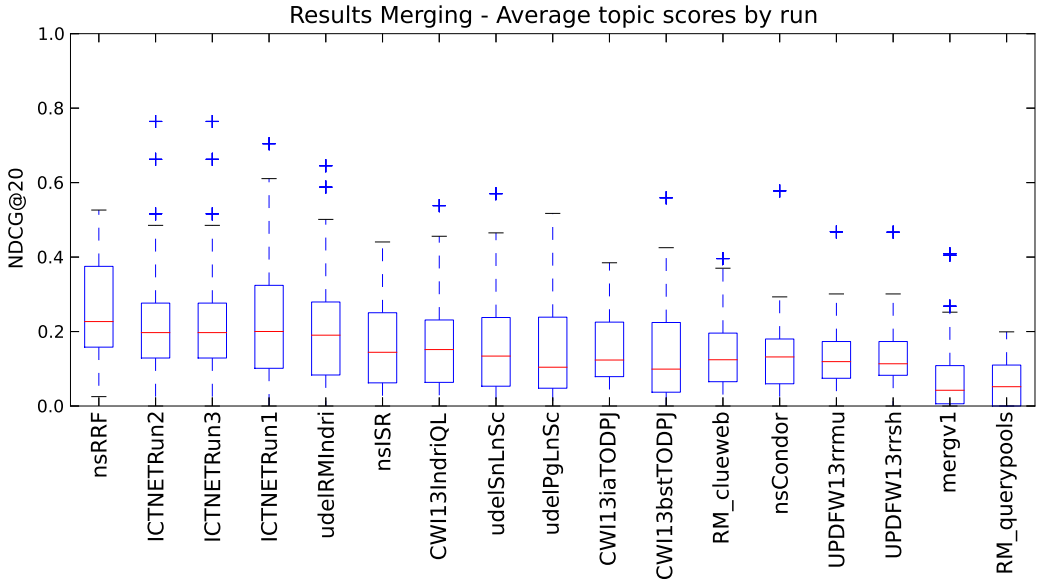

Figure 3: Comparison of runs on the Results Merging task.

the same URL. Two runs were submitted based on existing fusion approaches, Reciprocal rank fusion (`nsRRF`) and Condorcet Fuse (`nsCondor`). In addition, they submitted a run based on their own Inverse Square Rank approach (`nsISR`).

University of Padova (UPD)

UPD merged results in a round robin fashion [7]. The `UPDFW13rrsh` run was based on the ranking from the `UPDFW13sh` run, the `UPDFW13rrmu` run used the ranking obtained in the `UPDFW13mu` run.

Chinese Academy of Sciences (ICTNET)

They experimented with three different methods [13]. The `ICTNETRun1` run was created by scoring documents based on BM25 and combining the scores of each field (including URL, title, main content, headings) using a linear weighting method. The `ICTNETRun2` run also took the Google's pagerank score into account. `ICTNETRun3` filtered documents with a low score.

University of Delaware (udel)

Their baseline run (`udelRMIndri`) ranked the result documents using Indri. Next, they experimented with scoring the results by multiplying the natural logarithm of the resource scores with the normalized Indri-scores of the documents based on documents (`udelPgLnSc`) and snippets (`udelSnLnSc`)

Centrum Wiskunde & Informatica (CWI)

The baseline run of CWI [5] (`CWI13IndriQL`) scored documents based on their query likelihood. The `CWI13iaTODPJ` run was developed by assuming that by diversifying documents from different resources, it is more likely that at least one type of documents (resource) will satisfy the information need. The baseline run was reranked using a diversification algorithm (IA-select). They also experimented with boosting documents from reliable resources based on the resource selection scores `CWI13bstTODPJ`.

Indian Statistical Institute (isi_pal)

Their `mergv1` run was obtained by scoring documents based on he rank of the document in the results and the score of the resource (as calculated in the resource selection task) [18].

Organizers' baseline

The organizers' baseline runs used the static rankings from the corresponding size-based resource selection baselines (`RM_clueweb` and `RM_querypools`). The results of the top 5 ranked resources were combined using a round-robin merge.

5.3 Analysis

Most of the submitted and better performing runs for the results merging task make two unrealistic assumptions. Firstly, they asssume that for the given query all engine results are readily available. A more realistic scenario would be to first make a selection of a small number of promising engines, and to retrieve and rerank this set of results. Secondly, they assume that the result documents are readily available during search, whereas in a realistic scenario only the snippets would be available for real-time result merging. The few runs that do not make these assumptions and only use the top-ranked resources in combination with round-robin merging (e.g. from team UPD and the organizer's baseline runs) perform poorly in comparison to teams who indexed and searched the query search results from all engines.

As expected, not rewarding the retrieval of duplicate pages turns out to have a strong impact on the performance metrics. However, the nDCG@20 and nDCG@20(*) scores show a strong correlation (Pearson's r = 0.91, Kendall's tau = 0.79).

6. SUMMARY & OUTLOOK

The first edition of the Federated Web Search track attracted a total of 11 participants taking part in at least one of the two tasks: resource selection and result merging. The best performing resource selection runs were based on sample document indices in combination with standard document retrieval models. A baseline run which simply returned resources by its descending estimated size showed very competitive performance. The results merging runs were also dominated by standard retrieval approaches. Most of these runs are based on indices containing the retrieved documents from all search engines, which would be unrealistic in an online system.

Next year the same collection of search engines will be used with a new set of topics. In addition a new crawl of the samples will be made available – a comparison between the new and old samples could provide insight in the dynamics of the underlying resources and be useful for resource selection. The evaluation metrics for the tasks will be reviewed, for instance taking into account duplicate pages in resource selection. Next to the existing resource selection and results merging tasks, the track will feature a *vertical selection* task. In this task the systems have to rank the best vertical type for a query. What vertical types will be used is to be decided, but they will probably relate to the categories listed in Table 3.

7. ACKNOWLEDGMENTS

This work was funded by The Netherlands Organization for Scientific Research, NWO, grant 639.022.809, by the Folktales As Classifiable Texts (FACT) project in The Netherlands, by the Dutch national project COMMIT, and by Ghent University - iMinds in Belgium.

8. REFERENCES

[1] R. Aly, D. Hiemstra, D. Trieschnigg, and T. Demeester. Mirex and Taily at TREC 2013. In *Proceedings of the 22nd Text REtrieval Conference Proceedings (TREC)*, 2014.

[2] J. Arguello, F. Diaz, J. Callan, and B. Carterette. A methodology for evaluating aggregated search results. In *ECIR 2011*, pages 141–152, 2011.

[3] J. Arguello, F. Diaz, J. Callan, and J.-F. Crespo. Sources of evidence for vertical selection. In *SIGIR 2009*, pages 315–322, 2009.

[4] K. Balog. Collection and document language models for resource selection. In *Proceedings of the 22nd Text REtrieval Conference Proceedings (TREC)*, 2014.

[5] A. Bellogín, G. G. Gebremeskel, J. He, J. Lin, A. Said, T. Samar, A. P. de Vries, and J. B. P. Vuurens. CWI and TU Delft at TREC 2013: Contextual suggestion, federated web search, KBA, and web tracks. In *Proceedings of the 22nd Text REtrieval Conference Proceedings (TREC)*, 2014.

[6] A. Broder, M. Fontura, V. Josifovski, R. Kumar, R. Motwani, S. Nabar, R. Panigrahy, A. Tomkins, and Y. Xu. Estimating corpus size via queries. In *CIKM 2006*, pages 594–603, 2006.

[7] E. D. Buccio, I. Masiero, and M. Melucci. University of Padua at TREC 2013: federated web search track. In *Proceedings of the 22nd Text REtrieval Conference Proceedings (TREC)*, 2014.

[8] C. Burges, E. Renshaw, and M. Deeds. Learning to Rank using Gradient Descent. In *ICML 2005*, pages 89–96, 2005.

[9] J. Callen. Distributed information retrieval. In *Advances in Information Retrieval*, volume 7 of *The Information Retrieval Series*, chapter 5, pages 127–150. Springer, 2000.

[10] O. Chapelle, D. Metlzer, Y. Zhang, and P. Grinspan. Expected reciprocal rank for graded relevance. In *CIKM '09*, pages 621–630, 2009.

[11] M. S. Charikar. Similarity estimation techniques from rounding algorithms. In *STOC 2002*, pages 380–388, 2002.

[12] T. Demeester, D. Nguyen, D. Trieschnigg, C. Develder, and D. Hiemstra. What snippets say about pages in federated web search. In *AIRS 2012*, pages 250–261, 2012.

[13] F. Guan, Y. Xue, X. Yu, Y. Liu, and X. Cheng. ICTNET at federated web search track 2013. In *Proceedings of the 22nd Text REtrieval Conference Proceedings (TREC)*, 2014.

[14] D. Hawking and P. Thomas. Server selection methods in hybrid portal search. In *SIGIR 2005*, pages 75–82, 2005.

[15] J. Kekäläinen and K. Järvelin. Using Graded Relevance Assessments in IR Evaluation. *Journal of the American Society for Information Science and Technology*, 53(13):1120–1129, 2002.

[16] A. Mourão, F. Martins, and J. Magalhães. NovaSearch at TREC 2013 federated web search track: Experiments with rank fusion. In *Proceedings of the 22nd Text REtrieval Conference Proceedings (TREC)*, 2014.

[17] D. Nguyen, T. Demeester, D. Trieschnigg, and D. Hiemstra. Federated search in the wild: the combined power of over a hundred search engines. In *CIKM 2012*, pages 1874–1878, 2012.

[18] D. Pal and M. Mitra. ISI at the TREC 2013 federated task. In *Proceedings of the 22nd Text REtrieval Conference Proceedings (TREC)*, 2014.

[19] M. Shokouhi and L. Si. Federated search. *Foundations and Trends in Information Retrieval*, 5(1):1–102, 2011.

[20] D. Trieschnigg, K. Tjin-Kam-Jet, and D. Hiemstra. Ranking XPaths for extracting search result records. Technical Report TR-CTIT-12-08, Centre for Telematics and Information Technology, University of Twente, Enschede, 2012.

[21] D. Trieschnigg, K. Tjin-Kam-Jet, and D. Hiemstra. SearchResultFinder: Federated search made easy. In *SIGIR 2013*, pages 1113–1114, 2013.

APPENDIX

A. FEDWEB 2013 SEARCH ENGINES

ID	Name	Categories	ID	Name	Categories
e001	arXiv.org	Academic	e099	Bing News	News
e002	CCSB	Academic	e100	Chronicling America	News
e003	CERN Documents	Academic	e101	CNN	News
e004	CiteSeerX	Academic	e102	Forbes	News
e005	CiteULike	Academic	e103	Google News	News
e006	Economists Online	Academic	e104	JSOnline	News
e007	eScholarship	Academic	e106	Slate	News
e008	KFUPM ePrints	Academic	e107	The Guardian	News
e009	MPRA	Academic	e108	The Street	News
e010	MS Academic	Academic	e109	Washington post	News
e011	Nature	Academic	e110	HNSearch	News,Tech
e012	Organic Eprints	Academic	e111	Slashdot	News,Tech
e013	SpringerLink	Academic	e112	The Register	News,Tech
e014	U. Twente	Academic	e113	DeviantArt	Photo/Pictures
e015	UAB Digital	Academic	e114	Flickr	Photo/Pictures
e016	UQ eSpace	Academic	e115	Fotolia	Photo/Pictures
e017	PubMed	Academic,Health	e117	Getty Images	Photo/Pictures
e018	LastFM	Audio	e118	IconFinder	Photo/Pictures
e019	LYRICSnMUSIC	Audio	e119	NYPL Gallery	Photo/Pictures
e020	Comedy Central	Audio,Video	e120	OpenClipArt	Photo/Pictures
e021	Dailymotion	Audio,Video	e121	Photobucket	Photo/Pictures
e022	YouTube	Audio,Video	e122	Picasa	Photo/Pictures
e023	Google Blogs	Blogs	e123	Picsearch	Photo/Pictures
e024	LinkedIn Blog	Blogs	e124	Wikimedia	Photo/Pictures
e025	Tumblr	Blogs	e126	Funny or Die	Video,Photo/Pictures
e026	WordPress	Blogs	e127	4Shared	Audio,Video,Books,Photo/Pictures
e027	Columbus Library	Books	e128	AllExperts	Q&A
e028	Goodreads	Books	e129	Answers.com	Q&A
e029	Google Books	Books	e130	Chacha	Q&A
e030	NCSU Library	Books	e131	StackOverflow	Q&A
e032	IMDb	Encyclopedia	e132	Yahoo Answers	Q&A
e033	Wikibooks	Encyclopedia	e133	MetaOptimize	Academic,Q&A
e034	Wikipedia	Encyclopedia	e134	HowStuffWorks	Kids,Q&A
e036	Wikispecies	Encyclopedia	e135	AllRecipes	Recipes
e037	Wiktionary	Encyclopedia	e136	Cooking.com	Recipes
e038	E? Online	Entertainment	e137	Food Network	Recipes
e039	Entertainment Weekly	Entertainment	e138	Food.com	Recipes
e041	TMZ	Entertainment	e139	Meals.com	Recipes
e042	The Sun	Entertainment,Sports,News	e140	Amazon	Shopping
e043	Addicting games	Games	e141	ASOS	Shopping
e044	Amorgames	Games	e142	Craigslist	Shopping
e045	Crazy monkey games	Games	e143	eBay	Shopping
e047	GameNode	Games	e144	Overstock	Shopping
e048	Games.com	Games	e145	Powell's	Shopping
e049	Miniclip	Games	e146	Pronto	Shopping
e050	About.com	General	e147	Target	Shopping
e052	Ask	General	e148	Yahoo? Shopping	Shopping
e055	CMU ClueWeb	General	e152	Myspace	Social
e057	Gigablast	General	e153	Reddit	Social
e062	Baidu	General	e154	Tweepz	Social
e063	CDC	Health	e156	Cnet	Software
e064	Family Practice notebook	Health	e157	GitHub	Software
e065	Health Finder	Health	e158	SourceForge	Software
e066	HealthCentral	Health	e159	bleacher report	Sports
e067	HealthLine	Health	e160	ESPN	Sports
e068	Healthlinks.net	Health	e161	Fox Sports	Sports
e070	Mayo Clinic	Health	e162	NBA	Sports
e071	MedicineNet	Health	e163	NHL	Sports
e072	MedlinePlus	Health	e164	SB nation	Sports
e075	U. of Iowa hospitals and clinics	Health	e165	Sporting news	Sports
e076	WebMD	Health	e166	WWE	Sports
e077	Glassdoor	Jobs	e167	Ars Technica	Tech
e078	Jobsite	Jobs	e168	CNET	Tech
e079	LinkedIn Jobs	Jobs	e169	Technet	Tech
e080	Simply Hired	Jobs	e170	Technorati	Tech
e081	USAJobs	Jobs	e171	TechRepublic	Tech
e082	Comedy Central Jokes.com	Jokes	e172	TripAdvisor	Travel
e083	Kickass jokes	Jokes	e173	Wiki Travel	Travel
e085	Cartoon Network	Kids	e174	5min.com	Video
e086	Disney Family	Kids	e175	AOL Video	Video
e087	Factmonster	Kids	e176	Google Videos	Video
e088	Kidrex	Kids	e178	MeFeedia	Video
e089	KidsClicks?	Kids	e179	Metacafe	Video
e090	Nick jr	Kids	e181	National geographic	Video
e091	Nickelodeon	Kids	e182	Veoh	Video
e092	OER Commons	Kids	e184	Vimeo	Video
e093	Quintura Kids	Kids	e185	Yahoo Screen	Video
e095	Foursquare	Local	e200	BigWeb	General
e098	BBC	News			

B. FEDWEB 2013 SELECTED TEST QUERIES

ID	Query
7001	LHC collision publications
7003	Male circumcision
7004	z-machine
7007	Allen Ginsberg Howl review
7009	linkedin engineering
7018	audiobook Raymond e feist
7025	M/G/1 queue
7030	Lyrics Bangarang
7033	Porto
7034	sony vaio laptop
7039	import .csv excel
7040	vom fass gent
7042	bmw c1
7046	tuning fork
7047	Dewar flask
7056	ROADM
7067	used kindle
7068	Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition
7069	Eames chair
7075	zimerman chopin ballade
7076	Bouguereau
7080	lord of the rings hobbits theme
7084	Burn after reading review
7087	Jonathan Kreisberg discography
7089	varese ionisation
7090	eurovision 2012
7094	calculate inertia sphere
7096	touchpad scroll dell latitude
7097	best dum blonds
7099	lecture manova
7103	cystic fibrosis treatment
7109	best place to eat pho in new york
7115	pittsburgh steelers news
7124	yves saint laurent boots
7127	which cities surround long beach ca
7129	avg home edition
7132	massachusetts general hospital jobs
7145	why do cats purr
7209	crab dip appetizer
7258	swahili dishes
7348	map of the united states
7404	kobe bryant
7406	does my child have adhd
7407	kim kardashian pregnant
7415	most anticipated games of 2013
7465	xman sequel
7485	bachelor party jokes
7504	leiden schools
7505	ethnic myanmar
7506	I touch myself singer dead