FedWeb Greatest Hits: Presenting the New Test Collection for Federated Web Search

Thomas Demeester[1], Dolf Trieschnigg[2], Ke Zhou[3], Dong Nguyen[2], Djoerd Hiemstra[2]
[1] Ghent University - iMinds, Belgium
[2] University of Twente, The Netherlands
[3] Yahoo Labs London, United Kingdom
tdmeeste@intec.ugent.be, d.trieschnigg@utwente.nl,
kezhou@yahoo-inc.com, {d.nguyen, d.hiemstra}@utwente.nl

ABSTRACT

This paper presents 'FedWeb Greatest Hits', a large new test collection for research in web information retrieval. As a combination and extension of the datasets used in the TREC Federated Web Search Track, this collection opens up new research possibilities on federated web search challenges, as well as on various other problems.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

Keywords

Federated web search, test collection

1. INTRODUCTION

Research challenges for future search engines lie for an important part in *specializing* to domains, types of data, and kinds of applications. By specializing, search engines can improve their ranking functions dramatically. For instance, a web shop search engine benefits from ranking by price, a search engine for scientific papers by the number of citations, and a social network search engine by an item's number of shares, the time of the post, or the number of friends or followers. The need for specialized search engines is further driven by the wish to search different media types, such as images and videos, and the wish for new interaction paradigms, such as question answering and conversational search. The importance of specialization can be witnessed in evaluation initiatives such as the Text Retrieval Conference (TREC), which in 2015 includes a clinical decision support track, a track for personalized and localized search (the contextual suggestion track), and the microblog track. Actual users however, love a single entry point to all their information. And since the "one size fits all" approach, as argued above, is not the way forward, search engines need to support *distributed* or *aggregated* search [1, 5], that is to forward a query to (a selection of) search engines, and combine their results into one coherent search results page. Distributed and aggregated search is provided by major web search engines, like Google, Bing, Yandex, and Baidu, but it is also one of the main challenges of enterprise search [4].

The TREC Federated Web Search track (FedWeb) takes up the above challenges for a scenario in which an aggregated search engine distributes its queries to resources that are beyond the control of the search engine or enterprise, explicitly investigating the following tasks, which give rise to a number of additional challenges (see the track overview papers [2, 3]):

1: *Vertical selection*: select the most promising categories of resources for a given query,

2: *Resource selection*: select the most promising resources for a given query,

3: *Results merging*: combine the results of those resources into a single result list.

2. GREATEST HITS COLLECTION

The FedWeb track provides two collections (for FedWeb'13 and FedWeb'14) containing the top 10 search results from arround a 150 online search engines, in response to a large set of sample queries (for training) and a number of test topics (for testing). The search engine results pages consist of the HTML page, and for each hit the title, url, summary, and (when available) thumbnail image. Statistics of the collection are listed in Table 1. By combining both collections and extending them with previously unreleased data, the FedWeb Greatest Hits collection becomes a large and unique dataset that offers new analysis and research possibilities.[1]

The new collection includes the web documents for the FedWeb'14 official topics (which we did not provide before, to make the FedWeb'14 tasks more realistic), as well as for a set of extra topics, whereby the official and those extra topics form a total of 506 test topics, with full results gathered both in 2013 and 2014.

In addition, the collection includes full page judgments for a number of evaluation topics. For the 2013 data, we added a subset of double judgments, and a full set of snippet judgments, per topic judged by the same assessor that judged the

[1]FedWeb Greatest Hits will be made available for researchers from `http://fedwebgh.intec.ugent.be`.

WWW 2015 Companion, May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3473-0/15/05.
http://dx.doi.org/10.1145/2740908.2742755.

Table 1: Overview of the FedWeb Greatest Hits collection.

	Description	2013		2014	
	number of resources	157		149	
	total size (compressed)	277.7 GB	(105 GB)	506.3	(185 GB)
samples	queries	157 × 2,000		149 × 4,000	
	search result pages	301,652	21.8 GB	548,787	44.5 GB
	snippets	1.973,591	0.9 GB	3,616,551	2.0 GB
	documents	1,894,174	177.7 GB	3,482,602	376.6 GB
official topics	number of topics (judged)	200 (50)		275 (50 + 10)	
	search result pages	61,661	1.9 GB	39,321	2.8 GB
	snippets	143,298	0.07 GB	187.227	0.1 GB
	documents	136,079	16.7 GB	178.475	26.6 GB
	screenshots	131,812	27.6 GB	180.441	37.5 GB
	snippet judgments (double)	34,010 (6,253)		0	
	page judgments (double)	34,010 (7,027)		40,456	
	tuples with duplicate pages	4,601		12,946	
extra topics	number of topics	306		231	
	search result pages	46,062	3.0 GB	33,491	2.6 GB
	snippets	254,646	0.12 GB	190,862	0.11 GB
	documents	241,941	27.9 GB	181.828	22.6 GB
TREC FedWeb	example runs, evaluation scripts, qrels files for official tasks, documentation				

corresponding pages. Note that the snippets were judged on the *perceived* page relevance (estimated from the snippets), rather than the actual page relevance, but with the same relevance levels.

Research challenges. Various challenges can be taken on with the test collection. For example, the large number of resources and relatively large number (24) of verticals should allow refined modeling of the verticals, with the focus on vertical diversity. Another direction could be the design of user interfaces that blend the results of resources, or combine them in a novel way. Furthermore, the double judgments allow studying assessor agreement, and the snippet judgments facilitate research on snippet generation strategies, or on snippet quality vs. page relevance.

The collection also allows investigating the persistence of top 10 search results over time, given the overlap in the test topics, and a similar overlap in the sample queries (as the 4000 sample queries from 2014 are a superset of the 2000 sample queries from 2013).

We finish with just a small taste of the kind of information that becomes available with the aggregated dataset. By comparing the 2013 and the 2014 results in answer to the FedWeb'13 test topics, we find that an average of 32% of the 2013 results are still among the top 10 answers from the same resources, when receiving the same query a year later. There is however a wide spread over the different verticals. Where for the more dynamical verticals this fraction of 'persistent' results is much lower (e.g., 4% for *sports* or 5% for *blogs*), others appear more statical (e.g., 53% for *recipes*, or 58% for *encyclopedia*). We can also study the degree of relevance for these persistent results vs. those that disappeared in between the data crawls. We look at the fraction of pages that were judged to be highly relevant or key results. In some cases, this relevant fraction is considerably larger among the persistent than among the more 'volatile' results. For example, 11% vs. 7% for *encyclopedia*, and 19% vs. 5% for *social*. We assume that in both cases, the persistent results are of a higher quality or broader interest. For *kids*, meanwhile, older results seem less relevant (6% vs. 27%). For many other verticals (such as *general*, *video*, *pictures*), the difference in relevance distribution remains small.

3. CONCLUSIONS

We introduced the aggregated FedWeb Greatest Hits dataset, an enriched combination of the collections used for the TREC Federated Web Search track, aiming to encourage and equip researchers to tackle many new challenges currently arising in the field of web IR.

4. REFERENCES

[1] J. Callan. Distributed information retrieval. In W. B. Croft, editor, *Advances in Information Retrieval*, volume 7 of *The Information Retrieval Series*, pages 127–150. Springer US, 2000.

[2] T. Demeester, D. Trieschnigg, D. Nguyen, and D. Hiemstra. Overview of the TREC 2013 Federated Web Search Track. In *TREC*, 2013.

[3] T. Demeester, D. Trieschnigg, D. Nguyen, K. Zhou, and D. Hiemstra. Overview of the TREC 2014 Federated Web Search Track. In *TREC*, 2014.

[4] D. Hawking. Challenges in enterprise search. In *Proceedings of the 15th Australasian Database Conference*. Australian Computer Society, Inc., 2004.

[5] M. Lalmas. Aggregated search. In M. Melucci and R. Baeza-Yates, editors, *Advanced Topics in Information Retrieval*, volume 33 of *The Information Retrieval Series*, pages 109–123. Springer Berlin Heidelberg, 2011.