Resource Selection for Federated Search on the Web

Dong Nguyen[1], Thomas Demeester[2], Dolf Trieschnigg[1,3], Djoerd Hiemstra[1]
[1] University of Twente, The Netherlands
[2] iMinds, Ghent University, Belgium
[3] MyDataFactory, The Netherlands
{d.nguyen, r.b.trieschnigg, d.hiemstra}@utwente.nl, tdmeeste@intec.ugent.be

Abstract

A publicly available dataset for federated search reflecting a real web environment has long been absent, making it difficult for researchers to test the validity of their federated search algorithms for the web setting. We present several experiments and analyses on resource selection on the web using a recently released test collection containing the results from more than a hundred real search engines, ranging from large general web search engines such as Google, Bing and Yahoo to small domain-specific engines. First, we experiment with estimating the size of uncooperative search engines on the web using query based sampling and propose a new method using the ClueWeb09 dataset. We find the size estimates to be highly effective in resource selection. Second, we show that an optimized federated search system based on smaller web search engines can be an alternative to a system using large web search engines. Third, we provide an empirical comparison of several popular resource selection methods and find that these methods are not readily suitable for resource selection on the web. Challenges include the sparse resource descriptions and extremely skewed sizes of collections.

1 Introduction

Despite the obvious utility and necessity of using large general web search engines for navigating the internet, a large part of the web cannot be reached by these. The so-called *hidden* or *deep web*, is not easily accessible to web crawlers [19], and hence a large amount of content cannot be included in the web search indices. Usually these pages are dynamic and only accessible through a specific search engine. *Distributed information retrieval*, also called *federated search*, provides a solution to this problem. The queries are directly issued to search interfaces of collections so that crawling of these collections is not needed anymore. In this paper we focus on one specific task in federated search: *resource selection*, which aims at selecting suitable collections to forward a specific query to.

A large amount of research has been done in the field of federated search. However, an appropriate dataset reflecting a realistic web environment remained absent until recently. As a result, many proposed resource selection methods have been evaluated on distributed collections that were artificially composed by dividing TREC collections [22], according to several criteria (like topic or source). These test collections are very different from real search engines we find on the web, which have different retrieval methods, skewed sizes and heterogeneous content types (images, text, video etc). In addition, some of the proposed methods for resource selection make strong assumptions that may not hold on the web. For example, some of the proposed 'big document models' assume that all resources use a certain ranking method, while the 'small document models' are built on the idea that we are able to obtain a good ranking on a centralized sample index. As a result, it is not clear to what extent the proposed resource selection methods are suited for the web.

In this paper we use a newly released dataset presented in Nguyen et al. [16]. The dataset contains result pages from 108 actual web search engines, ranging from large general web search engines such as Google, Bing and Yahoo to small domain-specific engines.

We propose a new method for the estimation of collection sizes, based on samples obtained by query-based sampling, and assuming the ClueWeb09 web collection is a representative sample from the web. We also investigate the feasibility of web search without the use of general web search engines. For that, we compare an optimized federated search system based on smaller web search engines, with an optimized system using the large general web search engines. Finally, we present an empirical comparison of well-known resource selection methods, finding that they are not entirely suitable for this setting.

The remaining part of the paper is structured as follows. Section 2 discusses related work. Then, the dataset is described. In section 4, we present experiments on size estimation. In section 5 several benchmarking experiments are presented. We conclude with a discussion and future work.

2 Related work

A federated search system presents three challenges: *resource description*, obtaining a representation of the resource, *resource selection*, selecting suitable collections for a query and *resource merging*, creating a single ranked list from the returned results [4]. In this section we will focus on relevant work related to resource selection. We will also briefly review research on estimating the size of collections.

2.1 Resource selection

During the resource selection step, the broker selects a subset of the collections to send the query to. Several approaches have been explored for the resource selection problem. One line of research involves the *big document* models. These models represent each collection as one big document. Standard retrieval methods (or adaptations of them) are used to rank the big documents (collections) for a query. Examples are CORI [5], which uses a modified version of INQUERY, and language modeling approaches [27, 28]. Losing the document boundaries has several disadvantages. For example, the contents of the big document can be dominated by a few large documents.

To retain the document boundaries, *small document models* have been proposed. These have been found to be more effective than big document models. A query is issued on a centralized index of the document samples. The proposed methods vary in how the ranking or scores are used to select the collections. ReDDE [25] and CRCS [21] use the ranking of the documents in the sample index and a scaling factor to handle the varying sizes of collections. Contrary to ReDDE, the weight of a document in CRCS depends on the rank, while ReDDE uses only a cut off. GAVG [20] uses the score of the documents instead of the ranking, by calculating the geometric mean of the query likelihood of the top documents for each collection. These methods assume that the system is able to obtain a good ranking on a centralized sample index. However, when having multiple media types (e.g. images and text), this assumption may not hold.

Recently, methods have been proposed that frame the collection selection problem as a *classification* problem (e.g. [1], [13] and [14]). The main benefit of this approach is the ability to easily add different types of evidence for prediction and not only rely on content-based methods. For example, Arguello et al. [1] used logistic regression models and explored corpus features (such as CORI and ReDDE), query category features and click-through features. Hong et al. [13] also used logistic classification models, but added a term to account for the similarity between resources. This was motivated by the idea that a resource tends to be more relevant when similar resources have a high probability of being relevant.

Craswell et al. [8] experimented with resource selection for the web. To simulate a web environment where search engines have different retrieval algorithms, resources used different retrieval methods such as BM25 and Boolean ranking. Hawking and Thomas [11] evaluated resource selection methods in the GOV domain and proposed a hybrid approach combining distributed and central IR techniques. Although the work just discussed used actual web pages, the collections were relatively small compared to collections found on the web. For example, the largest collections were around 20 thousand documents [11], or a couple of thousand [8].

2.2 Size estimation

Many of the resource selection methods incorporate the size of a collection in their scoring method. However, when dealing with uncooperative resources, the actual size is not known and has to be estimated using the obtained samples. Methods to estimate the size of a collection assume random samples. Unfortunately, the samples from query based sampling are often not random, resulting in additional noise or challenges when estimating the size.

The Capture-Recapture [15] method uses the overlap of documents from two random samples to estimate the size, while Capture History [24] looks at overlapping documents when having a consecutive number of random samples. Sample - Resample [25] relies on document frequencies. Since many uncooperative resources do not provide (accurate) document frequencies, this method is often not suitable. Broder et al. [3] proposed two approaches to estimate the corpus size via query pools. The first relies on a uniform random sample of documents. The second approach is more practical and uses two (fairly) uncorrelated query pools and is related to the Capture-Recapture method.

3 Dataset description

In this section we briefly summarize the test collection presented in Nguyen et al. [16].

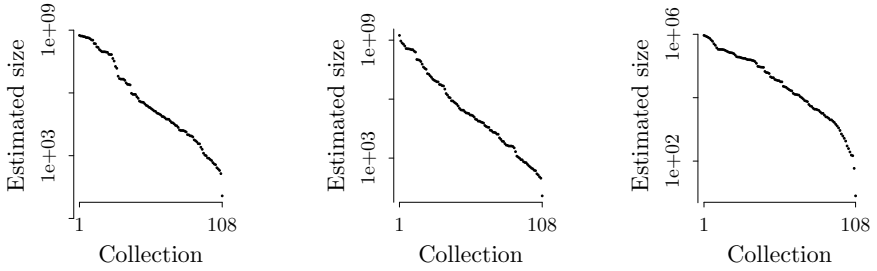
Motivation
The dataset was developed to reflect a realistic web environment, and designed to have the following properties: heterogeneous content, skewed sizes and relevance distributions, different retrieval algorithms and overlapping documents.

Search engines
The dataset contains 108 search engines, with a large variation in size and topic (e.g. sport, general web, news, academic, etc.). The dataset contains general web search engines such as Google and Bing, as well as smaller search engines focusing on a single domain or topic (for example arXiv.org, Photobucket, Starbucks and TechRepublic).

Sampling
Using query based sampling, samples were collected to inform the resource selection methods. The snippets as well as the documents of the search results were stored. Three methods were used to select queries for sampling: *Random*, *Top*, and *Zipf*. *Random* selects single terms randomly from the documents sampled so far. *Top* uses the most popular queries from a web search engine's query log [18]. *Zipf* uses single term queries taken evenly from the binned term distribution in ClueWeb09. For each query, the top 10 results from each engine were retrieved. Sampling was done by collecting the snippets of 197 queries and downloading the actual documents of the first 40 queries.

Topics
For testing, the dataset contains results of the search engines for the 50 topics of TREC Web Track 2010. These queries were developed to represent web search queries and are therefore very short. In addition, they were used to evaluate the Web Track diversity track and therefore many of them are ambiguous.

Relevance judgements
For the set of test queries, the relevance judgements for the top 10 results of each search engine were collected. Because many of the topics are ambiguous, annotators were instructed to judge according to the general information need indicated in the TREC Web Track 2010 full topic specification. For example for the query *titan* the general information need is *Find information on the Nissan Titan truck*. For the experiments in this paper, we use the relevance judgements based on the pages, but snippet judgments were collected as well. Six levels of page relevance were used, denoted in increasing order of relevance as *Junk*, *Non*, *Rel*, *HRel*, *Key*, and *Nav*, corresponding with the relevance levels from the Web



(a) ClueWeb (b) ClueWeb II (c) QueryPools - Zipf

Figure 1: Size distribution

TREC 2010 ad-hoc task [6]. For our analysis, we merged the *Junk* results into the *Non* category. For most of the experiments in this paper, a binary relevance is used where pages are considered relevant when scored on average better than *Rel*.

4 Size estimation

Estimating the size of a collection is a key step in many resource selection methods. In this section we discuss the used methods. In the next section we analyze the effect of the size estimation methods on resource selection. Note that the ratio of sizes among the collections is more important for resource selection experiments, than absolute numbers. We will reuse the samples obtained from query based sampling (to inform resource selection) to estimate the collection sizes. We explore two different methods to estimate the size of a collection. The first method involves a reference corpus (in our case ClueWeb09), the second method makes use of the number of overlapping documents in the samples.

4.1 Methods

ClueWeb I Our first method is built on the intuition to scale document frequencies using a reference corpus. We will use the ClueWeb09 dataset that represents a crawl of the web. For each query $q \in Q$ obtained using query based sampling, we calculate the size by dividing the number of results by its ClueWeb document frequency $df_{q,Clueweb}$ and scaling by the total ClueWeb size $|C_{Clueweb}|$. We then take the average of all queries to estimate the size $\hat{N}_i$ of collection C_i.

$$\hat{N}_i = \frac{1}{|Q|} \sum_{q \in Q} \frac{numResults_{q,i}}{df_{q,Clueweb}} \times |C_{Clueweb}|$$

The method is limited by the maximum number of results that is collected for the sample queries. Here, at most 10 results were collected per query, giving a maximum estimate of 10 times the ClueWeb size. However, a big advantage is that this maximum estimate is independent of the number of queries. The choice of reference corpus is important though, and choosing the ClueWeb09 dataset will estimate large sizes for collections that do well on most terms (have a broad coverage, just as the web). Very specific collections that only return results for a couple of queries, but are in fact very large, will have a low estimate using this method. This method is not robust against resources that always return the maximum number of results (regardless whether the results match the query or not), since it will then incorrectly give a very high estimate.

ClueWeb II The method using ClueWeb09 gives low estimates for very frequent terms such as *the* due to the limited number of results for each sample query (maximum of 10). Therefore, we also explore a

variant, that discards sample queries for which a resource returned 10 results (the maximum number) while the ClueWeb09 corpus has a higher document frequency for that term.

QueryPools Our second method is based on query pools and a modification of the method used by Broder et al. [3]. We divide the queries randomly into two sets to create two uncorrelated sets of queries (A and B), and use the document overlap $|D_{AB}|$ to estimate the size using $\frac{|D_A||D_B|}{|D_{AB}|}$ [15]. This method has two limitations. First, if there is no document overlap, the result is not well defined. In our estimates, we then take $|D_{AB}|$ to be 1. Second, the actual estimate is limited by the number of samples. When taking the minimum value of $|D_{AB}|$ to be 1, the maximum value is $|D_A||D_B|$. For example, to get an estimate of 600 mln, $|D_A|$ and $|D_B|$ should be 24.60 mln (assuming no overlap), an infeasible amount of samples.

4.2 Analysis

The *ClueWeb* method uses an external resource to estimate the size, while the *QueryPools* method uses the overlapping documents. The size distributions of the methods are plotted in Figure 1 and examples of estimated sizes are presented in Table 1. We find a very skewed size distribution using all methods. Because these search engines are real search engines on the web, their size is unknown. However, the results do suggest that the sizes are extremely skewed, much more than in previous test collections for federated search. Note that the exact size estimates differ by orders of magnitude when comparing both methods, with the ClueWeb method giving much higher estimates. The reason of this was outlined in the previous section.

ClueWeb I - Zipf		**ClueWeb II - Zipf**		**QueryPools - Zipf**	
Google	562 mln	Baidu	1,745 mln	Google	923k
Mamma.com	526 mln	Babylon	889 mln	Baidu	884k
Yahoo	521 mln	Bing	715 mln	Bing	843k
Baidu	495 mln	Google	621 mln	Ask	797k
IMDb	485 mln	Google Blogs	503 mln	Google Books	747k

Table 1: Largest resources

5 Resource Selection

We first outline the experimental setup. Then, we demonstrate the potential of a federated search system using the resource selection task. Next, we test the performance of well known resource selection methods on this dataset and discuss the results.

5.1 Experimental setup

Task For a given query q, the goal of resource selection is to select the top k best suited resources to search.

Evaluation We evaluate our methods using two metrics.

recall@k (also used by [25, 26, 21]) is the proportion of relevant documents in the top k collections of the resulting ranking Ω compared to the top k collections of the optimal ranking O (in this paper k is usually 5):

$$Recall = R_k = \frac{\sum_{i=1}^{k} \Omega_i}{\sum_{i=1}^{k} O_i}$$

precision@(10×k) Note that with the recall metric relevant documents returned by multiple collections are counted as relevant multiple times. To account for overlapping documents, we report the precision@(10×k), by assuming that each resource returns 10 results and calculating the precision after merging the results. By using the above cutoff, we overcome the problem of resource merging, we simply aggregate all results for the top k collections. Duplicate documents are only counted once (all others are considered irrelevant). This is similar to the evaluation method used by Shokouhi and Zobel [23], however they did use a resource merging method.

URL normalization Whether documents returned by different resources are in fact the same, is decided based on their URLs. Therefore, URLs were first normalized to a standard form, by deleting added strings such as search engine specific additions and query terms.

Binary relevance The page judgements were done on a six point scale. However, in this paper we will use binary relevance to evaluate the resource selection task. We have listed the proportions of page relevance categories in the annotations (without normalizing the URLs) in Table 2. Unless specified otherwise, we will consider a page to be relevant when on average it was ranked higher than *Rel*. We also sometimes use the very strict criterion higher than *Hrel*, such that less than 3 percent of the results is considered relevant.

Page relevance	Percentage
Junk/Non	82.841
Rel	8.756
HRel	5.533
Key	2.456
Nav	0.413

Table 2: Page judgements

5.2 Web search without the general web search engines?

In this section our goal is to analyze what performance can be obtained without the large web search engines in the context of resource selection [1]. We divide the resources into two sets:

1. *The general web search engines* (WSE's), containing 10 resources: Google, Bing, Yahoo, AOL, Ask, Mamma.com, Gigablast, Blekko, Baidu, Babylon.
2. *The remaining search engines* (indicated as non-WSE's), containing 98 resources. Many of them are very small, however some of these non-WSE's can still be very large, such as search engines of large sites such as YouTube, Flickr and Amazon, or specialized search engines like Google Blogs.

For each query, an oracle ranking for both resource sets was created based on the number of relevant documents. The recall is calculated by dividing by the optimal ranking with *all* search engines included. Results can be found in Table 3. Note that when evaluating at $k = 5$ we also select the most optimal resources for the WSE's (according to the number of relevant documents).

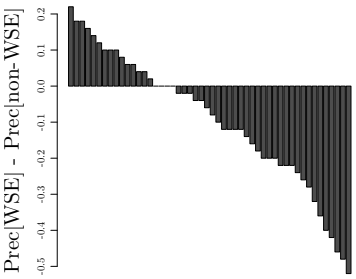
We find that for the recall metric the run using only general web search engines (WSE's) achieves higher performance, indicating that this system is able to achieve a higher total number of relevant documents. However, when we use the precision metric, which does take the unique number of results into account, the run without general web search engines performs better. We find that the WSE's return a lot of duplicate relevant documents.

Figure 2 shows a plot with the difference for precision per topic between the two sets. A positive value indicates that the oracle ranking using only WSE's performed better. Table 3 also shows a breakdown

[1] For our comparison between WSE's and non-WSE's, we only consider the potential number of relevant results; clearly other parameters play a role as well, e.g., the combined non-WSE's are most likely not capable of processing the huge workload of the WSE's.

	Only WSE		Non-WSE	
	Precision	**Recall**	**Precision**	**Recall**
k = 5	All: 0.248 Ambiguous: 0.198 Faceted: 0.307	All: 0.828 Ambiguous: 0.769 Faceted: 0.897	All: 0.343 Ambiguous: 0.304 Faceted: 0.388	All: 0.619 Ambiguous: 0.671 Faceted: 0.558
k = 10	All: 0.153 Ambiguous: 0.113 Faceted: 0.201	All: 0.742 Ambiguous: 0.674 Faceted: 0.822	All: 0.231 Ambiguous: 0.204 Faceted: 0.263	All: 0.514 Ambiguous: 0.568 Faceted: 0.449

Table 3: Oracle experiment



Precision of WSE vs. Non-WSE per top

Figure 2: Comparison oracles

for ambiguous versus faceted queries. We find that the WSE's do especially well on faceted queries. The oracle system using the other search engines does well on ambiguous queries when looking at recall, because in an oracle ranking the resources with the correct sense are ranked high.

An example query for which the oracle ranking using non-WSE's achieved better performance is the highly ambiguous query *avp* ('Find information about events sponsored by AVP, the Association of Volleyball Professionals.'). The top 5 of resources for this query are Myspace, WordPress, ESPN, Metacafe and Encyclopedia Britannica. An example of a query for which the oracle ranking using only WSE's was much better is *president of the united states* ('Find information about the office of President of the United States.'). The top 5 resources for this query are Gigablast, Ask, Yahoo, Mamma.com and Bing, all WSE's.

Next, we use a more strict criterion of relevance. We only consider pages to be relevant when they are on average judged higher than *HRel*. The results can be found in Table 4. Note that the average precision values are low, because the queries are highly ambiguous and we use a very strict criterion. We now observe that the general web search engines perform much better. For example, a system without general web search engines might miss a lot of relevant pages for navigational queries. Also, most of the test queries are general requests for information, hence better suited for the WSE's, as becomes clear further on.

	Only WSE		Non-WSE	
	Precision	**Recall**	**Precision**	**Recall**
k = 5	0.120	0.891	0.065	0.237
k = 10	0.070	0.844	0.033	0.187

Table 4: Oracle experiment - Better than HRel

When we use a less strict criterion of relevance (Table 5), we find the difference in precision between the oracles with WSE's and non-WSE's even bigger. Thus, the non-WSE's could be an alternative to

WSE's when taking a normal criterion on relevance. However, when the goal is to deliver high relevant results for example to answer navigational queries, the WSE's are still better suited.

	Only WSE		**Non-WSE**	
	Precision	**Recall**	**Precision**	**Recall**
k = 5	0.328	0.835	0.561	0.772
k = 10	0.217	0.735	0.437	0.684

Table 5: Oracle experiment - Rel or better

5.3 Resource selection experiments

As demonstrated before, an optimal resource selection system can be very powerful and a possible alternative to the general web search engines. We now test several well known resource selection methods for this task.

5.3.1 Preprocessing

Snippets were indexed using the text from the URL, title, and summary. The pages were indexed using the URL and text extracted from the document. Only text was extracted from HTML pages and PDF files. Some of the collections returned media files such as images (jpg, png, gif, svg) and sounds (ogg), for these files only the URL could be used for indexing. Terrier [17] was used for the retrieval experiments. No stopwords were removed, and words were stemmed using the Porter stemmer.

5.3.2 Baselines

We compare with the following baselines.

- *Popular* selects the top 5 most popular search engines in the US: Google, Bing, Yahoo, AOL and Ask [7].
- *Size Based 1* (SB1) ranks only those collections on size, that match at least one query term.
- *Size Based 2* (SB2) ranks all collections on size, resulting in the same ranking for each query.

5.3.3 Big document models

CORI is a big document model ([5], [4]) and an adaptation of the INQUERY method. Collections are represented by their terms and document frequencies and scored using:

$$p(w|C_i) = b + (1 - b) \times T \times I$$

with

$$T = \frac{df_{w,i}}{df_{w,i} + 50 + 150 \times \frac{cw_i}{avg_{cw}}}$$

$$I = \frac{\log(\frac{|C|+0.5}{cf_q})}{\log(|C| + 1.0)}$$

where $|C|$ is the number of collections, $df_{w,i}$ is the number of documents in C_i containing query term w, cw_i is the number of indexing terms in C_i, avg_{cw} is the average number of indexing terms in each collection and cf_q is the number of collections containing query term w. The free parameter b is usually

set to 0.4. We set it using cross fold validation. Collections are ranked by summing the beliefs of $P(w|C_i)$.

LM [27]. A big document is created for each collection, and using language modeling with Jelinek-Mercer smoothing the collections are ranked. G is the global language model by collapsing together all documents of all collections. λ is set using cross fold validation.

$$P(q|C_i) = \prod_{w \in q} \lambda P(w|C_i) + (1-\lambda) P(w|G)$$

5.3.4 Small document models

ReDDE (Relevant Document Distribution Estimation method) [25] estimates the number of relevant documents using the following equation:

$$Rel(C_i, q) = \frac{|C_i|}{|S_i|} \times \sum_{d \in S_i} P(rel|d, q)$$

Where $|C_i|$ is the estimated size of collection i, and $|S_i|$ is the number of samples for collection i. A query is issued to the sample index, the rank of a document d is then referred to as $sampleRank(d)$. ReDDE then estimates $centralRank(d)$, the rank of the document in the complete centralized index. A fixed probability k is then assigned to the top ranked documents in the centralized index (in our experiments $k = 1$).

$$P(rel|d, q) = k \text{ if } centralRank(d) < ratio * |C_{all}|$$

The *ratio* is usually set within a range of 0.002 to 0.005. We set it using cross validation. The centralized rank is estimated as follows:

$$centralRank(d) = \sum_{d', sampleRank(d') < sampleRank(d)} \frac{|C_{c'_d}|}{|S_{c'_d}|}$$

Where $|C_{c'_d}|$ is the size of the collection of document d'.

GAVG [20] scores a collection C_i using the geometric average query likelihood from C_i's top k documents. The query likelihood is calculated using Dirichlet smoothing. The smoothing parameter and k are determined using cross validation. If a collection has ranked less than k documents for the given query, the minimum query likelihood score is used for the missing documents.

$$GAVG(C_i, q) = \left(\prod_{d \in \text{top } k \text{ from } C_i} P(q|d) \right)^{\frac{1}{k}}$$

CRCS (Central rank based collection selection) [21] scores a collection according to:

$$CRCS(C_i, q) = \frac{|C_i|}{|C^{max}| \times |S_i|} \times \sum_{d \in S_i} R(d))$$

with $|C_i|$ the estimated size of collection C_i, $|C^{max}|$ the size of the largest collection and $|S_i|$ the size of the sample. $R(d)$ is nonzero when document d is ranked in the top k documents of the sample index. In contrary to ReDDE, the relevance of a document $R(d)$ depends on its $sampleRank(d)$. Note that the $sampleRank(d)$ is the same as used in ReDDE. Two variations are proposed to calculate $R(d)$.

$$\text{Linear: } R(d) = k - sampleRank(d)$$

$$\text{Exponential: } R(d) = \alpha \times exp(-\beta \times sampleRank(d))$$

The parameters k, α and β are set using cross fold validation. Preliminary experiments showed no significant difference between the linear and exponential version, therefore we have only used the linear version in our final experiments.

5.4 Results

In the following paragraphs, we will give the results for the measures and experiments introduced in Section 4 and 5.

Size
The results are presented in Figure 3. We find that methods that explicitly take the size of the collection into account achieve a much higher performance. Even a simple baseline, where all resources that match at least one of the query terms are ranked on size (SB1), achieves a high performance compared to the other methods. There are several possible explanations. The queries might be biased towards larger collections, the resource selection methods might not be suitable for this setting and/or the samplings are not sufficient. We analyze this further in the next section.

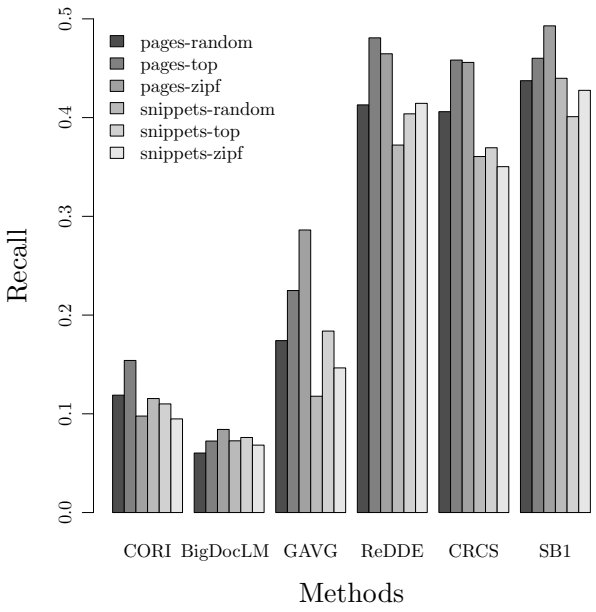


Figure 3: Results

Methods comparison
In addition we can make the following observations: First, small document models perform better than big document models, for example when comparing GAVG with CORI and LM. Second, from the big document models, CORI is consistently better than LM. Third, ReDDE is consistently better than CRCS. Fourth, for methods that take size into account, sampling using pages achieves higher performance. For the other methods, the results between the different sampling methods are mixed.

Performance on non-WSE's
An additional analysis was done by filtering out the 10 general web search engines from the results. For all methods, performance decreased. However, we found the same observations as just outlined with the results containing all search engines.

Baselines
The results for the baselines (see Section 5.3.2) that do not depend on the sampling are presented in Table 6. First, we find that always selecting the 5 major search engines results in a relatively high

performance. This confirms that the queries from the Web track are highly biased towards a general web search scenario. In addition, a simple ranking based only on size (using the ClueWeb Zipf method) also performs well (indicated as Size Based 2). Using only Google, the average recall is 0.574 and the precision is 0.392 for the top 10 results.

Baseline	**Recall at 5**	**Precision at 5**
Popular	0.646	0.2144
Size Based 2	0.477	0.1588

Table 6: Baseline results - Better than Rel

Overall, the performance of the resource selection methods in combination with the used query samplings is relatively low. In the next section we analyze the effect of size and the limited samplings in more depth.

5.5 Discussion

5.5.1 Importance of size

We found that the ranking methods that incorporate size perform relatively well. In this section we provide further analyses on the effect of size. First we ask:

To what extent do the resource selection methods correlate with a size based ranking?

We calculate the correlations between the ranking of the resource selection methods and a sized based ranking (Table 7) using Kendalls Tau. Because only a partial set of collections is ranked, we only calculate the correlation over the resources that are considered by the ranking method. We report the average correlation of a method by averaging across queries and sampling methods. We find that the methods that take size explicitly into account have a high correlation, while the other methods have a small negative correlation.

Method	τ
Big document LM	-0.085
CORI	-0.017
GAVG	-0.013
ReDDE	0.595
CRCS	0.487

Table 7: Correlation with size based ranking

To illustrate why methods such as ReDDE have such a high correlation with size, we provide an example calculation of the scaling factor that is used: $\frac{|C_i|}{|S_i|}$. For a resource like Google, the ClueWeb method estimated a size of $|C_i| = 561,965,753$. The sampling size for *Zipf* samples using pages was $|S_i| = 389$, resulting in a scaling factor of 561,965,753/389 = 1,444,642. Compare this for example to Fox Sports, which gets a scaling factor of only 16,027/139 = 115. It is questionable whether such a high influence of size is desired.

Next, we investigate the following question:

How is performance influenced by the choice of size estimation method?

Table 8 shows a comparison of the used methods by measuring the recall and precision at 5. Both methods perform similarly, although using *top* samples results in a low performance, probably because these queries are far from random.

Table 9 shows the results using only non-WSE's.

Method	Recall at 5	Precision at 5
Clueweb I - Zipf	0.493	0.1704
Clueweb II - Zipf	0.467	0.1856
QueryPool - Top	0.117	0.0644
QueryPool - Zipf	0.448	0.1932

Table 8: SB1 - Zipf pages sampling

Method	Recall at 5	Precision at 5
Clueweb I - Zipf	0.150	0.091
Clueweb II - Zipf	0.172	0.105
QueryPool - Top	0.110	0.067
QueryPool - Zipf	0.184	0.114

Table 9: SB1 - Zipf pages sampling - non-WSE

One striking observation of the results discussed above is that the size based rankings perform so well. We will now look closer into this observation and we ask:

> When is a purely size based ranking effective, and when are more sophisticated resource selection methods effective?

We make the following hypothesis: on queries for which the ranking quality of the documents in the sample index is low (for example when the sample index has not enough samples matching the query), a ranking only based on size will perform better than resource sampling methods.

In our analysis we will analyze the SB1 ranking using size estimated by the ClueWeb size estimation method based on page-based Zipf sampling. We will compare with the other resource selection methods, using an index based on Zipf sampling pages.

We make pairs of the SB1 ranking with the other resource selection methods, and calculate the difference between recall@5. A positive difference indicates that the SB1 ranking performed better than the other method.

Next, we calculate the correlation between the difference and an estimate of the ranking quality for the queries. We use the Simplified Clarity Score (SCS) as proposed in He and Ounis [12], which is based on query clarity [9] . He and Ounis [12] found a strong positive correlation between SCS and AP for short queries. A higher value for SCS therefore corresponds with a better ranking quality.

$$SCS = \sum_{Q} P_{ml}(w|Q) log_2 \frac{P_{ml}(w|Q)}{P_{coll}(w)}$$

with

$$P_{ml}(w|Q) = \frac{qtf}{ql}$$

Where qtf is the query term frequency, and ql is the query length. $P_{coll}(w)$ is the probability of term w in the collection. The found correlations are shown in Table 10.

For every method, we find a low but consistently negative correlation. This shows that the SB1 method has a tendency to be more effective compared to other methods for queries where the ranking quality is (predicted to be) low. We therefore expect a possible future strategy for resource selection methods that have to deal with sparse samples is to back off to a size based ranking for queries that are considered difficult.

Method	Correlation
Big document LM	-0.140
CORI	-0.153
GAVG	-0.067
ReDDE	-0.142
CRCS	-0.128

Table 10: Pearson correlation with SCS and performance of SB1 compared with other methods

Sampling method	Avg of total	Avg of relevant
Zipf-pages	0.51	0.72
Zipf-snippets	0.47	0.61
Top-pages	0.55	0.76
Top-snippets	0.46	0.57
Random-pages	0.58	0.73
Random-snippets	0.50	0.59

Table 11: Average fraction of resources considered for Web TREC 2010 queries

5.5.2 Is sampling enough?

In this section we analyze how performance is influenced by the amount of samplings.

How big is the ranking problem?
How many resource match at least one of the query terms such that it is considered in the ranking? The results are presented in Table 11 (avg of total). On average, the number of resources to rank is around 50-60 out of 108, and using pages the number is consistently higher than using snippets.

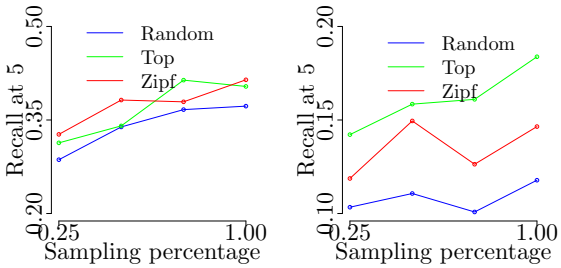
How many resources are we missing?
Next we analyze how many resources we are not even considering for ranking because of the limited sampling. In particular, for each query we divide the number of relevant resources (a resource that has at least one relevant document) that match at least one of the query terms by the total number of relevant resources for that query. We then average across queries. Results are found in Table 11 (avg of relevant). For example, for a given query 60 resources are considered for ranking, but only 20 of them are relevant. If the total number of relevant resources is 50, it means that we only consider $20/50 = 0.4$ of the relevant resources for ranking. We find that on average only 50-75% of the relevant resources are considered for ranking, indicating that many relevant resources are missed because of the limited samplings.

Varying the amount of samples
Now, we are interested in how the amount of sampling influences performance. We vary the amount of sampling available, by taking proportions of the obtained snippet samples by randomly selecting 0.25, 0.50 and 0.75 of the queries. Results are plotted in Figure 4. We analyze the performance of two methods ReDDE and GAVG. We observe that ReDDE shows an increasing performance as more data is available. The trend with GAVG is more noisy, although the pattern still suggests an increasing line.

Overall, we can conclude that the limited amount of sampling did have a large influence on the performance. We expect, and suggest for future work, that more samplings are desirable for this setting.

6 Dataset robustness

We will use the experiments from the previous section to assess the robustness of the test collection for research in resource selection for federated search. We will experiment with two measures, the Cronbach



(a) ReDDE (b) GAVG

Figure 4: Effect of sampling size

Alpha coefficient and the correlation between rankings based on subsets.

6.1 Cronbach Alpha

The Cronbach Alpha is a coefficient for measuring the reliability of tests. Recently it has been used to measure the reliability of IR datasets (e.g. [2], [10]). The coefficient gives an indication to what degree the questions test a single construct. The coefficient will be high, when there is a high correlation between the queries as well as with the end score (for example MAP or average recall@5). The coefficient is calculated as follows:

$$\hat{\alpha} = \frac{k}{k-1}(1 - \frac{\sum_i \hat{\sigma}_i^2}{\hat{\sigma}_x^2})$$

Where k is the number of queries, $\hat{\sigma}_i^2$ is the estimated variance for query i and $\hat{\sigma}_x^2$ is the estimated variance of the total scores. The test is simple to calculate, but it does not take into account individual variance of items. The test requires as input existing runs, we use the runs performed for our empirical comparison of the resource selection methods and use as scores the recall@5. We find a value of 0.98, which indicates a very high degree of reliability.

6.2 Subset topics

Next we analyze how many topics are needed to obtain a good ranking of the systems. We select random subsets from the topics, starting at size 5 and increasing with step size 5 to the full set of topics. For each subset size, we take multiple random samples (10) to calculate an average based on recall at 5. We calculate the correlation using Kendalls Tau between the ranking obtained on the subset with the ranking using the full dataset. The found correlations are shown in Figure 5.

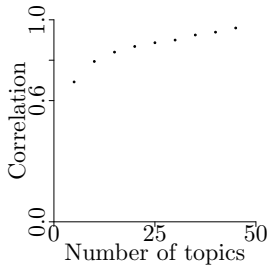


Figure 5: Correlation with full ranking

Around 30 topics, a correlation higher or equal to 0.75 is obtained. An correlation higher than 0.86 is achieved after 40 topics or more.

7 Conclusions

A dataset for federated search on the web, reflecting a real web environment, has long been absent. Recently, a new test collection was released [16], containing the results from more than a hundred real search engines, ranging from large general web search engines such as Google and Bing, to small domain-specific engines. In this paper analyses and experiments on resource selection were presented using this new dataset.

We demonstrated that federated search has a lot of potential. Without the large general web search engines, an optimal resource selection algorithm would be able to achieve comparable or better performance when looking at the unique number of relevant results. However, general web search engines are still useful for delivering highly relevant pages for example to answer navigational queries.

We experimented with several methods to estimate the size of uncooperative resources on the web. We used a method based on the number of overlapping documents and also introduced a method that uses an external reference corpus (ClueWeb09) to estimate the sizes. The sizes were found to be very effective for the resource selection methods.

Existing resource selection methods were found not to be readily suitable for the web setting. We found that due to the extremely skewed size distribution of the dataset, the size of the collections had an enormous influence on the performance. Resource selection methods that take the estimated size of a collection into account performed well compared to methods that do not include size in their ranking. Also, methods that take the size of a collection into account correlate highly with just a size based ranking. We found that size based rankings tend to be effective compared to more sophisticated methods when the (predicted) ranking quality in the sample index is low.

We suggest several future research directions for resource selection for the web. First, is the large influence of size in current resource selection methods desirable when the sizes are extremely skewed? Second, more research on size estimation is encouraged, especially in combination with query based sampling methods. And third, how to handle sparse resource descriptions and how many samples are desirable? Results suggest that sampling more documents is desirable for obtaining a more complete sample based index and the typical number of 300-500 documents is not sufficient for this setting to obtain a high performance.

8 Acknowledgements

This research was supported by the Dutch national program COMMIT and the Folktales as Classifiable Texts (FACT) project, which is part of the CATCH programme funded by the Netherlands Organisation for Scientific Research (NWO).

References

[1] J. Arguello, J. Callan, and F. Diaz. Classification-based resource selection. In *CIKM 2009*, pages 1277–1286. ACM, 2009.

[2] D. Bodoff and P. Li. Test theory for assessing ir test collections. In *SIGIR 2007*, pages 367–374. ACM, 2007.

[3] 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*, CIKM '06, pages 594–603. ACM, 2006.

[4] J. Callan. *Advances in Information Retrieval.*, chapter Distributed information retrieval, pages 127–150. Kluwer Academic Publishers, 2000.

[5] J. P. Callan, Z. Lu, and W. B. Croft. Searching distributed collections with inference networks. In *SIGIR 1995*, pages 21–28. ACM, 1995.

[6] C. L. A. Clarke, N. Craswell, I. Soboroff, and G. V. Cormack. Overview of the trec 2010 web track. In *TREC*, 2010.

[7] comScore. comScore Releases February 2012 U.S. Search Engine Rankings, 2012. [Online; accessed 19-April-2012].

[8] N. Craswell, P. Bailey, and D. Hawking. Server selection on the world wide web. In *Proceedings of the fifth ACM conference on Digital libraries*, DL '00, pages 37–46. ACM, 2000.

[9] S. Cronen-Townsend, Y. Zhou, and W. B. Croft. Predicting query performance. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '02, pages 299–306, New York, NY, USA, 2002. ACM.

[10] A. Harpale, Y. Yang, S. Gopal, D. He, and Z. Yue. Citedata: a new multi-faceted dataset for evaluating personalized search performance. In *CIKM 2010*, pages 549–558. ACM, 2010.

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

[12] B. He and I. Ounis. Inferring query performance using pre-retrieval predictors. In A. Apostolico and M. Melucci, editors, *String Processing and Information Retrieval*, volume 3246 of *Lecture Notes in Computer Science*, pages 229–248. Springer Berlin / Heidelberg, 2004.

[13] D. Hong, L. Si, P. Bracke, M. Witt, and T. Juchcinski. A joint probabilistic classification model for resource selection. In *SIGIR 2010*, pages 98–105. ACM, 2010.

[14] J. Kim and W. B. Croft. Ranking using multiple document types in desktop search. In *SIGIR 2010*, pages 50–57. ACM, 2010.

[15] K.-L. Liu, C. Yu, and W. Meng. Discovering the representative of a search engine. In *CIKM 2002*, pages 652–654. ACM, 2002.

[16] D. Nguyen, T. Demeester, D. Trieschnigg, and D. Hiemstra. Federated search in the wild. In *CIKM 2012*, 2012.

[17] I. Ounis, G. Amati, V. Plachouras, B. He, C. Macdonald, and C. Lioma. Terrier: A High Performance and Scalable Information Retrieval Platform. In *Proceedings of ACM SIGIR'06 Workshop on Open Source Information Retrieval (OSIR 2006)*, 2006.

[18] G. Pass, A. Chowdhury, and C. Torgeson. A picture of search. In *InfoScale 2006*. ACM, 2006.

[19] S. Raghavan and H. Garcia-Molina. Crawling the hidden web. In *VLDB 2001*. ACM, 2001.

[20] J. Seo and W. B. Croft. Blog site search using resource selection. In *CIKM 2008*, pages 1053–1062. ACM, 2008.

[21] M. Shokouhi. Central-rank-based collection selection in uncooperative distributed information retrieval. In *ECIR 2007*, pages 160–172. Springer-Verlag, 2007.

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

[23] M. Shokouhi and J. Zobel. Federated text retrieval from uncooperative overlapped collections. In *SIGIR 2007*, pages 495–502. ACM, 2007.

[24] M. Shokouhi, J. Zobel, F. Scholer, and S. M. M. Tahaghoghi. Capturing collection size for distributed non-cooperative retrieval. In *SIGIR 2006*, pages 316–323. ACM, 2006.

[25] L. Si and J. Callan. Relevant document distribution estimation method for resource selection. In *SIGIR 2003*, pages 298–305. ACM, 2003.

[26] L. Si and J. Callan. Unified utility maximization framework for resource selection. In *CIKM 2004*, pages 32–41. ACM, 2004.

[27] L. Si, R. Jin, J. Callan, and P. Ogilvie. A language modeling framework for resource selection and results merging. In *CIKM 2002*, pages 391–397. ACM, 2002.

[28] J. Xu and W. B. Croft. Cluster-based language models for distributed retrieval. In *SIGIR 1999*, pages 254–261. ACM, 1999.