Search Result Diversification in Resource Selection for Federated Search

Dzung Hong
Department of Computer Science
Purdue University
250 N. University Street
West Lafayette, IN 47907, USA
dthong@cs.purdue.edu

Luo Si
Department of Computer Science
Purdue University
250 N. University Street
West Lafayette, IN 47907, USA
lsi@cs.purdue.edu

ABSTRACT

Prior research in resource selection for federated search mainly focused on selecting a small number of information sources that are most relevant to a user query. However, result novelty and diversification are largely unexplored, which does not reflect the various kinds of information needs of users in real world applications.

This paper proposes two general approaches to model both result relevance and diversification in selecting sources, in order to provide more comprehensive coverage of multiple aspects of a user query. The first approach focuses on diversifying the document ranking on a centralized sample database before selecting information sources under the framework of Relevant Document Distribution Estimation (ReDDE). The second approach first evaluates the relevance of information sources with respect to each aspect of the query, and then ranks the sources based on the novelty and relevance that they offer. Both approaches can be applied with a wide range of existing resource selection algorithms such as ReDDE, CRCS, CORI and Big Document. Moreover, this paper proposes a learning based approach to combine multiple resource selection algorithms for result diversification, which can further improve the performance. We propose a set of new metrics for resource selection in federated search to evaluate the diversification performance of different approaches. To our best knowledge, this is the first piece of work that addresses the problem of search result diversification in federated search. The effectiveness of the proposed approaches has been demonstrated by an extensive set of experiments on the federated search testbed of the Clueweb dataset.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

SIGIR'13, July 28–August 1, 2013, Dublin, Ireland.

Keywords

Federated Search, Resource Selection, Diversification

1. INTRODUCTION

Federated search, also known as distributed information retrieval [28, 4, 14], focuses on searching information distributed across multiple information sources such as local repositories or verticals. There are three major sub-problems in federated search: resource representation obtains information about contents and other key properties of each individual information source, resource selection selects a small number of most useful sources given a user query, and result merging integrates individual ranked lists from selected sources into a single final list. A large body of research has been conducted for resource selection in federated search [4, 28]. However, little is known about selecting a set of sources that balances relevance and novelty. This substantially limits the usability of federated search in many applications.

On the other side, search result diversification has been studied extensively in ad hoc search in order to offer more coverage for ambiguous and multifaceted queries. In several occasions, users' intents in their queries may not be expressed explicitly. For example, an ambiguous query such as "Jaguar" may refer to an animal or a car model; or a multifaceted query such as "Batman" may refer to a name of a movie, a comic character, or the comic itself. Search result diversification circumvents this problem by explicitly or implicitly considering probable aspects of the query and presenting the search results in a way that is easier for users to find the needed information. Since 2009, the TREC Web track has incorporated diversification in the evaluation of the Web track [9]. Several evaluation metrics have been developed in ad hoc search to measure the effectiveness of different approaches of search result diversification.

Search result diversification in federated search may not be as simple as diversifying the final ranked list obtained from the selected sources. In a federated environment where documents of a same source cover similar topics, selecting a set of sources that balances relevance and novelty becomes crucial. As for the example of the query "Jaguar" above, if the sources related to "Jaguar" as a car dominate the resource selection result, it will be much harder to obtain a diversified ranked list in the end.

This paper proposes new approaches in diversifying results of resource selection in federated search. To the best of our knowledge, this is the first study that tackles the issue. First, a set of new metrics is designed for measuring result diver-

sification in resource selection. The metrics can incorporate any diversity measure that has been developed in ad hoc search, including the intent-aware expected reciprocal rank (ERR) [7], α-nDCG@k [13] and MAP-IA [1]. Second, two general approaches are proposed for diversifying resource selection. The first approach extends the ReDDE framework and utilizes a ranked list of documents on the centralized sample database. By reranking the sample documents with respect to result diversification, a better set of sources can be obtained in term of relevance and aspect coverage. The second approach offers a different view. Instead of reranking the sample documents based on their relevance to the query aspects, the information sources are reranked in a similar process. This can be done by estimating the relevance of each information source with respect to different aspects of the query by any existing resource selection algorithm. Furthermore, a learning based classification approach is proposed to combine multiple resource selection algorithms for a better estimation of source relevance with respect to query aspects. With some training data, the learning based classification approach can improve the effectiveness of resource selection for search result diversification.

An extensive set of experiments has been conducted with the federated search testbed of the Clueweb dataset to evaluate the proposed research. In many experimental settings, the new approaches can successfully improve result diversification over traditional approaches that only consider document relevance. In particular, the new approaches provide superior performance on two test levels: source-level results of resource selection and document-level results of the final ranked list of federated document retrieval. Finally, the learning based approach, which combines results from multiple resource selection algorithms, outperforms each individual algorithm in result diversification.

The rest of the paper is organized as follows. Section 2 offers the literature review on both resource selection in federated search and result diversification in ad hoc search. Section 3 discusses our proposed metric to measure the diversity of selected sources. Section 4 proposes the two result diversification approaches for resource selection in federated search. Section 5 presents the learning based classification approach. The new proposed research is examined by an extensive set of experiments in Sections 6 and 7. Section 8 concludes our work and points out some potential research directions in the future.

2. RELATED WORK

Considerable research have been conducted for all three sub-problems of federated search as resource representation, resource selection and result merging [4, 28, 15]. This section provides a discussion of prior research on resource selection, as well as a brief review on resource representation and result merging. We also discuss some popular ad hoc search algorithms for search result diversification.

Resource representation is the first step of federated search for obtaining important properties of distributed information sources such as content and size statistics. Query-based sampling [4] is a common approach as it obtains sample documents from available sources with randomly generated queries. The sample documents obtained in this process can be placed together in a *centralized sample database*.

Resource selection selects a small number of most relevant information sources for a user query. Some early resource selection algorithms such as CORI [4], CVV [37] and KL [36] treat each source as a big document and derive useful statistics to rank available sources with respect to a user query. However, these algorithms have limitation of losing the boundaries of individual documents, and thus may underestimate a big source with many relevant documents. Topic modeling has been proposed by recent work to overcome this limitation [3].

Other resource selection algorithms such as ReDDE [32], DTF [17], CRCS [27] and SUSHI [33] step away from the big document assumption by modeling individual documents of a source. Several selection algorithms in this category rely on the centralized sample database to build a ranked list of sample documents for a user query, and then assign a relevance score to available sources based on the scores of their sample documents in the list. Different algorithms use different methods for aggregating document contribution to available sources. Recent work by Markov et al. follows a similar approach, but attempts to minimize uncertainty in the centralized sample database by sampling different queries, retrieval systems, or rankings [22].

Learning based models have also been proposed for resource selection. They treat resource selection in federated search [20], or vertical search [2] as a classification problem. In particular, given a set of training queries and some relevance judgment, a classification model can learn to predict the relevance of an information source. In some experiments, the classification approaches have been shown to provide more accurate resource selection results than traditional algorithms without the training process.

Result merging is the last step in federated search, which merges documents returned by selected sources into a single ranked list. Modern methods such as SSL [31] and SAFE [30] both attempt to merge documents by approximating the centralize retrieval results in different ways.

Existing research in federated search have not explored an important issue of result novelty and diversification, which limits their abilities in representing the various information needs of users. The research work in [29] estimates the degree of document overlap among available sources, but its focus is only on duplicate documents and does not directly address the diversification problem related with multiple aspects of user queries. Other research work in [25] and [39] address diversification in aggregated search, which is similar to federated search, but operates in cooperative environments. Most importantly, those work do not target diversification in selecting relevant verticals (or sources) directly.

On the other hand, result diversification has been a popular research topic in ad hoc search. Its goal is to make a trade-off between relevance and novelty in ranking documents [38, 5, 1, 6, 26]. In order to achieve the desired effect of covering sufficient aspects of a user query (so that user will likely find the sought information), diversification algorithms target on discovering novel aspects that have not been covered in the ranked list, and reducing the redundancy information shared between multiple documents.

The earlier generation of diversification algorithms do not explicitly consider multiple aspects of a query [5, 38]. Instead, they build the ranked list from top to bottom, and make a choice of whether to include a document based on its similarity with existing documents in the list. More recent diversification algorithms directly incorporate query aspects into consideration. Agrawal et al. proposed the use of tax-

onomy to classify query aspects, in order to discover novel and redundant information [1]. Carterette and Chandar directly optimized the ranked list with respect to evaluation measures based on diversity [6]. Santos et al. proposed the eXplitcit Query Aspect Diversification (xQuAD) method, which estimates the surplus information that a document can add to a ranked list, using the probability of relevance with respect to all aspects of the query [26]. Recently, Dang and Croft proposed a different view on diversification by preserving the proportionality of document presence with respect to each aspect of the query [16]. Their proposed algorithm PM-2 has proven to achieve superior performance over several other algorithms of the same category. Other recent work on search result diversification include a combination of implicit and explicit topic representations [19], personalized diversification [34], and explicit relevance model [35].

3. DIVERSIFICATION METRICS IN RESOURCE SELECTION FOR FEDERATED SEARCH

There exists several standard diversification metrics for ad hoc search. However, no evaluation metric has been developed to compare the diversification results of different resource selection algorithms in federated search. This section proposes a new metric in order to fill that gap. Our metric is based on the popular R-metric [4] for resource selection in federated search. It is calculated as the ratio between the number of relevant documents contained in sources selected by a particular algorithm, over the number of relevance documents in sources selected by an ideal algorithm. In particular, the R-metric is defined as:

$$R_k = \frac{\sum_{i=1}^{k} E_i}{\sum_{i=1}^{k} B_i}$$

where E_i denotes the number of relevant documents of the i-th source according to the ranking E by a particular resource selection algorithm, and B_i denotes the same quantity with respect to the optimal ranking B. In this case, the optimal ranking B should order sources by the true number of their relevant documents. We adopt the idea to our new generic metric, which is called R-based diversification metric, as follows.

$$R_{\mathcal{M}}(\mathbf{S}) = \frac{\mathcal{M}(\text{optimal ranking of documents in } \mathbf{S})}{\mathcal{M}(\text{optimal ranking of documents in all sources})}$$

where $\mathbf{S}$ is the set of selected sources for comparison and $\mathcal{M}$ is a diversity metric of a ranked list of documents such as ERR-IA, α-nDCG, Prec-IA, S-Recall and NRBP. The optimal ranking of documents in $\mathbf{S}$ is the list that achieves the best score with respect to metric $\mathcal{M}$. For most of the aforementioned metrics, finding the optimal ranked list is an NP-hard problem, but this can be acceptably approximated by a greedy algorithm (i.e., repeatedly select the next document that maximizes the metric given the current ranking; cf. [8, 13]). The intuition of the proposed metric is that, if we can select a minimal set of sources that contains enough diversified documents to reach the optimal measure, then the R-based diversification metric is maximized to 1. Otherwise, it gives us an estimation of how far our selected sources are from the optimal ones.

Like R-metric, the proposed R-based diversification metric is independent of the retrieval algorithm utilized by each source. The R-based diversification metric returns 1 if all available sources are selected. In general, for comparison between different resource selection algorithms, the maximum number of sources is determined beforehand.

4. TWO APPROACHES FOR DIVERSIFICATION IN RESOURCE SELECTION

This section proposes two approaches for selecting diversified information sources. The first approach extends the ReDDE framework by ranking sample documents with consideration to diversity. The second approach estimates the source relevance to each aspect of the query and ranks the sources based on the estimations of their aspect relevance.

4.1 Diversification on Sample Documents under ReDDE framework

We first describe the Relevant Document Distribution Estimation (ReDDE) framework [32] for ranking sources based on the centralized sample database. In this framework, a given query is issued to the centralized sample database to retrieve a ranked list of sample documents. ReDDE makes an assumption that each sample document in this list represents a number of (unseen) documents from the source that it belongs to. Based on that, a source score is calculated by aggregating all contribution from its sample documents. The amount of contribution is scaled up depending on the source's size. Original ReDDE assigns a constant score for all documents on the top part of the returned list, and multiplies that constant with the ratio of the estimated source size over the sample size. The obtained quantity is then used for aggregating source scores. The CRCS resource selection algorithm [27] follows the same approach, but varies the amount of contribution of each document by an exponential decay function, as documents further down the ranked list have less contribution to its source.

ReDDE and many other algorithms of the same family such as CRCS and SUSHI [33] utilize a ranking on the centralized sample database to estimate the relevance of available sources. For algorithms of this family, they mostly vary in the way of defining a utility function for each document in the list. Having said that, the original ranked list from the centralized sample database plays an important role. Non-diversification algorithms target on building a centralized ranked list that covers as many relevant documents to the query as possible. In many cases, this may not pay enough attention to sources that cover multiple query aspects. For diversification purpose, we should be careful when selecting a source that mainly contains documents relevant to an aspect that has been covered before.

The above observation suggests a way to achieve good search diversification results by constructing a ranked list that covers several aspects of the query. We call this approach Diversification approach based on sample Documents (DivD). Instead of building a centralized ranked list that focuses only on relevance, we construct a ranked list that offers more diversity. The goal is to reduce the contribution of a document (on behalf of its source) that may be relevant to the query, but offers less novelty in the overall ranking.

This approach can combine a wide range of resource selection algorithms with any diversification algorithm that has been developed before, for instance, PM-2 [16] and xQuAD [26], which were mentioned in Section 2. A typical example

Algorithm 1 Diversification Approach based on Sample Documents using ReDDE and PM-2

1: Initialize scores of all sources to 0
2: Rank documents of the centralized sample database by an effective retrieval algorithm, e.g. Indri
3: Rerank that list by PM-2
4: **for** each document on top of the ranked list **do**
5: Add a constant score c to the source containing the document
6: **end for**
7: **return** the ranked list of all sources based on their scores

Algorithm 2 Source-based Diversification using ReDDE and PM-2

1: Rank all sources using standard ReDDE
2: **for** each aspect q_i of query q and each source s_j **do**
3: Estimate $P(s_j|q_i)$, the probability of relevance of source s_j with respect to q_i using the ReDDE algorithm
4: **end for**
5: Rerank the list obtained in step 1 using PM-2 algorithm with the estimated $P(s_j|q_i)$ as inputs
6: **return** the ranked list of all sources based on their diversification scores

of combining the standard ReDDE and PM-2 is shown in Algorithm 1.

Some resource selection algorithms utilize the relevance score of each document in the centralized sample ranked list for ranking sources (e.g. ReDDE.top [2]). In the diversification approach that has been discussed so far, this relevance score can be replaced by the diversity score given by a diversification algorithm for ad hoc search. There are different interpretations about those diversity scores, depending on the assumptions made by the diversification algorithms. In our experiments, diversity scores can be used with most resource selection algorithms effectively.

4.2 Diversification Approach based on Source-level Estimation

The second approach follows a different strategy than the first one. As presented in the previous section, the diversification approach based on sample documents works directly on the ranked list of sample documents, which is not a natural component for resource selection algorithms similar to CORI. Indeed, several resource selection algorithms utilize summary statistics of a source to estimate its relevance to a query. It is not straightforward to apply a diversification method based on individual sample documents for those kinds of algorithms.

This paper proposes another diversification approach for resource selection that operates at the source level. More specifically, many existing diversification algorithms for ad hoc search rank documents by estimating their relevance with respect to each aspect of the query, and then harvesting this information in order to produce a ranked list that balances multiple query aspects. It is possible to design a similar process that uses the *estimated source relevance* with respect to each aspect of the query. More specifically, instead of building a diversified list of documents for a given query as in ad hoc search, we can build a diversified list of available sources in a similar manner. All estimations with respect to the documents can be replaced by estimations with respect to the sources. This resource selection approach for results diversification is called Diversification approach based on Source-level estimation (DivS).

An important step in DivS is therefore to compute the probability of relevance of a source with respect to a query aspect. Many existing resource selection algorithms are able to provide such information. CORI can directly provide a relevance estimation based on the big document assumption. ReDDE and CRCS do not provide direct estimations, but it is possible to use their source scores aggregated from sample documents for such a purpose. An example of the diversification approach based on source-level estimation using ReDDE and PM-2 algorithms is presented in Algorithm 2.

Compared with diversification approach based on sample documents, the diversification approach based on source-level estimation can work with a wider range of resource selection algorithms. On the other hand, it requires multiple runs of a resource selection algorithm for all different aspects of a query, which is more time consuming than the former approach. In real-world applications, it is possible to design a parallel solution for multiple resource selection runs to speed up the process.

5. A CLASSIFICATION APPROACH FOR COMBINING DIVERSIFICATION RESULTS

The two diversification approaches based on sample documents and source-level estimation both utilize a specific resource selection algorithm and a diversification algorithm for ad hoc search. It is possible to combine the results of multiple resource selection algorithms for search result diversification, which may provide better results by modeling complementary results from different algorithms. This section proposes a learning based classification approach. With some training information, this method can learn how to combine evidence supporting available sources from different resource selection algorithms for result diversification.

In particular, we adapt the classification-based approach that has been used for both vertical search [2] and federated search [20]. For collecting training information in learning the classification model, this approach generates a pseudo-relevant judgment of a source given a query by counting the number of relevant documents that the source contains. If the number is higher than some threshold value τ, the source is considered to be relevant. In this paper, the training dataset for result diversification consists of multiple instances, each of them represents a pair of a source and a query aspect. A source is considered relevant to a query aspect if it contains at least one document relevant to that aspect of the query. For each pair of a source and a query aspect, we utilize the source score information from the following resource selection algorithms as features:

- **ReDDE.top** [2] This is quite similar to the original ReDDE algorithm, but uses the relevance scores of documents in calculating source scores, instead of a step function as in traditional ReDDE. ReDDE.top has

been chosen since it provides more consistent results than ReDDE in our dataset.

- **CRCS** [27] with exponential decay function for estimating probability of relevance of sample documents. This method has been reported by several studies to be better than the linear decay version [33].
- **Big Document** This is a traditional approach that collapses all sample documents of a source into a big document, which represents this source and is used for resource selection. Our retrieval algorithm for this approach is Indri [23].
- **CORI** [4] CORI is another type of big document selection approach with a different tf.idf weighting scheme. It shares the same assumption that considers each source as a big document of its sample documents.

More specifically, the learning based approach attempts to naturally integrate evidence from two different views of resource selection algorithms, one based on big document assumption (Big Document and CORI), and the other based on aggregated information of sample documents (ReDDE.top and CRCS). In the experimental section, we will provide more analysis and some examples about why this combination strategy may outperform each individual method.

All the features provided by the above algorithms are normalized for each query in order to achieve more consistency. Given the training dataset, it is possible to design a learning method that estimates $P(s|q_k)$ for a new query aspect q_k. We choose logistic regression model as it has been shown to be among the best in many practical applications such as text categorization [18].

Let s_i^j be the binary variable that indicates the relevance of the i-th source to the query q_j, i.e. $s_i^j = 1$ indicates relevance, and $s_i^j = 0$ indicates otherwise. Let $\mathbf{f}_i^j$ be the vector of all features returned by the resource selection algorithms mentioned above. We can then represent the relevance probability of source s_i given $\mathbf{f}_i^j$ by a sigmoid function σ:

$$P(s_i^j = 1|q_j) = \frac{exp(\mathbf{f}_i^j \cdot \mathbf{w})}{1 + exp(\mathbf{f}_i^j \cdot \mathbf{w})} = \sigma(\mathbf{f}_i^j \cdot \mathbf{w})$$

where $\mathbf{w}$ denotes the combination weight vector.

Learning the combination weight $\mathbf{w}$ can be conducted by maximizing the log-likelihood function using the iterative reweighted least squares method. The learned parameter can be then used to estimate the relevance probability $P(s|q_k)$ for any particular aspect of a new user query. This probability becomes inputs for the diversification approach based on source-level estimation to rank the sources. Our hypothesis is that, if the learning model can provide a more accurate estimation than those produced by a single resource selection algorithm, we can expect the learning based approach to generate more accurate results. This approach is denoted as LR-DivS, as it applies logistic regression within the diversification approach based on source-level estimation.

6. EXPERIMENTAL METHODOLOGY

Dataset.

The experiments in the paper have been conducted on the federated search testbed of the Clueweb English dataset.

Table 1: Document statistics of the federated search testbed based on Clueweb English

# of sources	total # of docs	min # of docs	max # of docs	avg # of docs
2,780	151,161,188	48	3,417,805	54,364.27

The Clueweb dataset[1] is a large collection of web pages that has been used in several official tasks in the Text REtrieval Conference (TREC) tracks. Furthermore, the available information of queries with multiple aspects and the corresponding relevance judgment has enabled the evaluation of search result diversification. The federated search testbed derived from Clueweb is publicly available[2] as an attempt to offer a large and realistic testing environment for federated search. This collection contains 2,780 information sources and about 151 million documents, which is much larger than most other testbeds that have been previously used in federated search. Information sources are created by collecting web documents of the same domains (e.g. *blogspot.com, about.com*), and the Wikipedia documents are clustered into 100 collections using the K-means clustering algorithm with two passes/iterations. More statistics information of this dataset is given in Table 1. It is noticed that Clueweb has also been used in distributed environment, albeit in a different problem setting [21].

We are aware of a recent dataset that has been proposed for federated search in web search environment [24]. However, this dataset does not provide relevance judgments on multiple aspects of a query, and thus does not fully support experiments of this research at the moment.

Each information source in the testbed has been assigned a retrieval algorithm, which was chosen from a set of algorithms such as Inquery, Language Model and tf.idf in a round-robin manner. This strategy simulates the behavior that information sources in real world applications may use different types of retrieval algorithms. In order to build the centralized sample database, 300 documents have been sampled from each source via the source's specific retrieval algorithm. The Indri retrieval algorithm [23] was used in all retrieval processes on the centralized sample database.

The queries used in our experiments consist of 148 queries from the TREC Web Track 2009[9], 2010[10], and 2011[11].

Diversification Algorithm.

Diversity by Proportionality [16] (PM-2) and eXplicit Query Aspect Diversification (xQuAD) [26], which are among the state-of-the-art diversification algorithms, have been examined in our study. In particular, we notice that the performance of PM-2 tends to be better than xQuAD in most of the metrics, thus we only report our results based on PM-2. In our implementation of this algorithm, we have chosen to rerank the top $K = 500$ documents in the centralized sample database, as well as in the final step of diversifying results from the centralized complete database (i.e., all documents in available sources). The parameter λ in PM-2 is set to be 0.5 in all settings.

[1] http://boston.lti.cs.cmu.edu/Data/clueweb09
[2] http://www.cs.purdue.edu/homes/dthong/clueweb/

An important component of PM-2 and many other search result diversification algorithms is the estimation of $P(d|q_i)$, which is the relevance score of a document d with respect to a particular aspect q_i of a query. These aspects are usually not available in real-world applications. We follow the work of [16, 26] to report our results on two scenarios: all the aspects of a query are provided; or we can retrieve the aspects from a commercial search engine such as Google or Bing. For the first scenario, we use subtopics that come with TREC queries as aspects. For the second scenario, we send the original query to the search engine (Google in our experiments), and adopt its suggestions as query aspects. The second set contains 144 queries, as 6 original queries do not have search engine's suggestions at the moment of our work. We name the first scenario "Given Subtopics" and the second "Suggestions" in the results presented in Section 7.

Resource Selection Algorithms.

A set of commonly used resource selection algorithms as described in section 5 has been utilized in the new research of result diversification in resource selection. They consist of ReDDE.top, CRCS with exponential decay function, Big Document and CORI.

With CRCS, the top 500 sample documents returned by the centralized sample database are considered. The exponential decay function of CRCS makes it more stable on the Clueweb collection, which exhibits a highly skewed distribution of source sizes. On the other side, ReDDE.top is more sensitive to noise in such an environment, as a not-so-relevant sample document from a really big source may result in too much bias in favor of that source and affects the final ranked list. Therefore, for ReDDE.top, we set the number of top sample documents for each query to be smaller than CRCS's, chosen from the set $\{50, 100, 150, \cdots\}$. We report the results using top 50 documents for ReDDE.top as it provides the most consistent performance.

To name the different methods, we use a prefix "D" for a diversification resource selection algorithm to indicate an approach based on sample documents, and a prefix "S" to indicate an approach based on source-level estimation. As discussed before, the Big Document and CORI algorithms have only the S versions. For all methods reported in the next section, we select up to 10 sources for each query.

Training and Testing.

All the proposed resource selection algorithms for search result diversification do not need any training data except the final approach of combining multiple resource selection algorithms in a learning model (LR-DivS). Therefore, for the algorithms other than LR-DivS, we report the results on all 148 queries from TREC Web Track 2009-2011. For LR-DivS, since it requires a training dataset, we use queries with TREC id less than or equal to 75 for training, and the rest for testing. We also report the results on the testing set of all other algorithms for comparison. For the scenario with suggested aspects of user queries, since there is no corresponding relevance judgment, we train our model using the provided query subtopics/aspects and their corresponding relevance judgments. However, in the testing phase, the model is applied with features derived from the suggested aspects, i.e., we estimate $P(s|q_i)$ where q_i is a query aspect suggested by a commercial search engine. Finally, we evaluate all approaches with respect to TREC's provided subtopics/aspects. This strategy is consistent with the evaluation process in TREC Web Track [11].

Evaluation Metric.

The proposed new research has been evaluated at two levels: source selection and federated document retrieval.

- **Diversification results with R-based diversification metric for source selection**: We evaluate the resource selection results using the R-based diversification metric described in Section 3. In particular, five popular metrics in result diversification such as ERR-IA[7], α-nDCG[13], NRBP[12], P-IA (intent aware precision [1]) and S-Recall (subtopic recall, for the number of subtopics/aspects covered by top documents) have been adopted with the R metric in resource selection and used in the experiments.
- **Diversification results for federated document retrieval**: To make the evaluation independent from a specific result merging algorithm, the Indri algorithm is used to perform document retrieval on the centralized complete database. Only documents from the selected sources for each query have been retained, which is consistent with prior research in [32]. This ranked list of documents is reranked again using the PM-2 algorithm. We then evaluate the final ranked list by the five metrics mentioned above as ERR-IA, α-nDCG, NRBP, P-IA and S-Recall. All of these metrics are computed at the top 20 documents, which is consistent with the official TREC evaluation of search result diversification for ad hoc search [11] and consistent with the commonly used high-precision metric in federated document retrieval.

7. EXPERIMENTAL RESULTS

An extensive set of experiments has been conducted for evaluating several approaches proposed in this paper, which are: the approach based on sample documents (DivD), the approach based on source-level estimation (DivS) and the learning based approach (LR-DivS). We conduct experiments on two levels for different purposes: source-level results for resource selection and document-level results for federated document retrieval. More specifically, the first subsection compares the performance of a diversification approach with standard resource selection algorithms. The second set of experiments in 7.2 compares multiple resource selection algorithms adapting DivD and DivS approaches. The third set of experiments in 7.3 demonstrates the advantage of the learning based approach (LR-DivS) over approaches with a single resource selection algorithm. The last set of experiments in 7.4 compares the document-level diversification results across all proposed approaches.

7.1 Diversification versus Standard Resource Selection Algorithms

This subsection compares the performance of two standard resource selection algorithms ReDDE.top and CRCS with their diversification counterparts at source selection level. In particular, we choose to study the first diversification approach based on sample documents (DivD) in this subsection as they are more related with standard ReDDE.top and CRCS algorithms, while more results of both diversification approaches will be presented shortly.

Table 2 shows the performance using all the R-based diversification metrics described in the previous section. Without diversification, it can be observed that the standard CRCS significantly outperforms standard ReDDE.top in all metrics. This may be attributed to the fact that CRCS generally selects more relevant sources to the query, which leads to a wider range of aspects being covered. The advantage of CRCS may come from using the exponential decay function for document utility, which tends to be better than using document score as utility in ReDDE.top. When the document-based diversification approach is applied, it further increases the performance of the standard algorithms: D-ReDDE.top significantly outperforms ReDDE.top in its capacity of selecting diversified sources. As for CRCS, its diversification version (i.e., D-CRCS) is also consistently better than the standard CRCS algorithm. The same observation can be seen in both scenarios when the query aspects are given, or suggested by Web search.

7.2 Diversification with Different Resource Selection Algorithms

This subsection studies the performance of the two proposed resource selection approaches (i.e., DivD and DivS) using several resource selection algorithms, including ReDDE-top, CRCS, Big Document and CORI. The results are presented in Table 3. In all settings, the standard Big Document and standard CORI algorithms are outperformed by the other methods. Furthermore, both S-Big Document and S-CORI, which are under the same assumption of collapsing sample documents within a source, are inferior to S-ReDDE.top and S-CRCS. These results indicate that ReDDE-top and CRCS tend to be more effective in resource selection for result diversification than Big Document and CORI, which is consistent with previous research in federated search for resource selection without diversification.

Both DivD and DivS approaches produce comparable results when applied to ReDDE.top and CRCS. The D-CRCS version based on sample documents is better than its counterpart based on source-level estimation (i.e., S-CRCS), whereas the contrary is observed for ReDDE.top. In case of ReDDE.top, the difference is significant, which may be explained by the fact that when the original ranked list is short (only 50 for ReDDE.top), it is more difficult for the diversification algorithm to find a document that covers many query aspects, rather than finding a source that covers many aspects.

The results using the provided query aspects and suggested ones reveal an interesting observation. We notice that the performance of the two settings are quite comparable with little difference. One possible explanation is that, in federated environment, it may not need a perfect set of query aspects for selecting a diversified set of information sources, as sources are already somehow divided by different types of semantic topics. Since our goal is to select sources that can cover as many query aspects as possible, the resource selection algorithms can do a reasonably good work as long as the suggestions of query aspects provide some meaningful interpretations of different aspects of the query.

7.3 Classification Approach for Combining Diversification Results

This subsection compares the performance of the learning based classification approach with all other diversification approaches mentioned above. Since the classification approach requires a set of training queries and does the testing on another set, we also report the results of all previous methods on the testing queries for comparison. Table 4 presents the results. The performance of the standard resource selection algorithms and their diversification counterparts are better on the set of testing queries than on the set of all queries, due to the particular division of training and testing queries. The comparison between standard algorithms' performance and those of diversification approaches on the testing queries raises similar observations as mentioned in the previous subsections 7.1 and 7.2.

It can be seen that the classification approach provides the best performance over all metrics. It can be attributed to the fact that the learning based classification approach can harness the advantage of different algorithms, and combine them in an effective way. A typical example from our training set is the query "Obama family tree" with its provided subtopic "Find the TIME magazine photo essay Barack Obama's Family Tree". For ReDDE.top and CRCS, it is almost impossible to find a sample document containing all the keywords of the subtopic/aspect, if such a document does not exist in the sample database. On the other hand, Big Document and CORI can provide some useful hints for selecting sources by looking at all sample documents from each source as whole. For example, several different sample documents of a source may contain various parts of the query. This is particularly useful for sources that cover a wide range of topics, for instance, Wikipedia. Another example is the query "Volvo" with its provided subtopic "Find a Volvo dealer". Big Document and CORI can give hint to the sources that contain the words "Volvo" and "dealer" from different sample documents. For those sources, the classification approach can utilize the complementary results for improving search results diversification.

7.4 Diversification Results of Federated Document Retrieval

This subsection compares the search diversification results in document-level of federated document retrieval. Given the top ten sources selected by the proposed algorithms for each query, the final ranked list generated from the sources is evaluated. One set of results compares algorithms without training over all queries. The other set of results is included for the test queries to compare LR-DivS with other algorithms. We omit the results of Big Document and CORI on the first set due to space limitation.

Table 5 provides the results on all queries. The standard ReDDE.top algorithm falls behind the other models in its diversification capacity. Among all the methods without training, D-CRCS consistently outperforms the other methods, which is consistent with its performance on the R-based diversification metrics in the source-level.

Table 6 provides the results on the test queries. Again, the same trend with algorithms without training can be observed. When some training information is available to learn how to combine multiple evidence, the classification approach LR-DivS consistently provides the best document-level diversification performance among all models.

8. CONCLUSION AND FUTURE WORK

Resource selection is an important research problem in federated search for selecting a small number of relevance sources for a given query. Various algorithms have been pro-

Table 2: R-based diversification metric of resource selection on all 148 queries. Symbols † and ‡ indicate statistical significance under paired t-test with respect to ReDDE.top baseline and CRCS baseline ($p < 0.05$).

			R-ERR	R-α-nDCG	R-NRBP	R-P-IA	R-S-Recall
Given Subtopics	Baseline	ReDDE.top	0.549	0.529	0.566	0.406	0.545
		CRCS	$0.677^{\dagger}$	$0.652^{\dagger}$	$0.698^{\dagger}$	$0.484^{\dagger}$	$0.658^{\dagger}$
	DDiv	D-ReDDE.top	$0.624^{\dagger}$	$0.598^{\dagger}$	$0.645^{\dagger}$	$0.438^{\dagger}$	$0.612^{\dagger}$
		D-CRCS	$\mathbf{0.699}^{\dagger}$	$\mathbf{0.672}^{\dagger}$	$\mathbf{0.721}^{\dagger}$	$\mathbf{0.499}^{\dagger\ddagger}$	$\mathbf{0.674}^{\dagger}$
Suggestions	Baseline	ReDDE.top	0.542	0.522	0.558	0.396	0.541
		CRCS	$0.673^{\dagger}$	$0.649^{\dagger}$	$0.694^{\dagger}$	$0.477^{\dagger}$	$0.659^{\dagger}$
	DDiv	D-ReDDE.top	$0.614^{\dagger}$	$0.589^{\dagger}$	$0.635^{\dagger}$	$0.434^{\dagger}$	$0.604^{\dagger}$
		D-CRCS	$\mathbf{0.698}^{\dagger}$	$\mathbf{0.671}^{\dagger}$	$\mathbf{0.720}^{\dagger}$	$\mathbf{0.491}^{\dagger\ddagger}$	$\mathbf{0.677}^{\dagger}$

Table 3: R-based diversification metric on resource selection on multiple diversification approaches. Letters r, c, R, C, B, O indicate statistical significance under paired t-test to D-ReDDE.top, D-CRCS, S-ReDDE.top, S-CRCS, S-BigDoc, and S-CORI respectively ($p < 0.05$).

			R-ERR	R-α-nDCG	R-NRBP	R-P-IA	R-S-Recall
Given Subtopics	DDiv	D-ReDDE.top	0.624_{B}	0.598_{BO}	0.645_{B}	0.438_{BO}	0.612_{B}
		D-CRCS	$\mathbf{0.699}_{rBO}$	$\mathbf{0.672}_{rBO}$	$\mathbf{0.721}_{rBO}$	$\mathbf{0.499}_{rCBO}$	$\mathbf{0.674}_{rBO}$
	SDiv	S-ReDDE.top	0.680_{rBO}	0.656_{rBO}	0.701_{rBO}	0.496_{rCBO}	0.664_{rBO}
		S-CRCS	0.675_{rBO}	0.649_{rBO}	0.698_{rBO}	0.462_{rBO}	0.664_{rBO}
		S-BigDoc	0.490	0.468	0.506	0.346	0.486
		S-CORI	0.569_{B}	0.541_{B}	0.591_{B}	0.351_{B}	0.561_{B}
Suggestions	DDiv	D-ReDDE.top	0.614_{B}	0.589_{B}	0.635_{B}	0.434_{BO}	0.604_{B}
		D-CRCS	$\mathbf{0.698}_{rBO}$	$\mathbf{0.671}_{rBO}$	$\mathbf{0.720}_{rBO}$	$\mathbf{0.491}_{rCBO}$	$\mathbf{0.677}_{rBO}$
	SDiv	S-ReDDE.top	0.677_{rBO}	0.652_{rBO}	0.698_{rBO}	0.485_{rCBO}	0.666_{rBO}
		S-CRCS	0.673_{rBO}	0.647_{rBO}	0.695_{rBO}	0.458_{BO}	0.665_{rBO}
		S-BigDoc	0.490	0.469	0.508	0.342	0.490
		S-CORI	0.566_{B}	0.538_{B}	0.587_{B}	0.347_{B}	0.562_{B}

Table 4: R-based diversification metric of resource selection on multiple diversification approaches on test queries. Symbols $\dagger, \ddagger, r, c, R, C, B, O$ indicate significant improvement under paired t-test to ReDDE.top, CRCS, D-ReDDE.top, D-CRCS, S-ReDDE.top, S-CRCS, S-BigDoc, and S-CORI respectively. ($p < 0.05$).

			R-ERR	R-α-nDCG	R-NRBP	R-P-IA	R-S-Recall
Given Subtopics	Baseline	ReDDE.top	0.618	0.604	0.630	0.461	0.635
		CRCS	$0.769^{\dagger}_{rBO}$	$0.751^{\dagger}_{rBO}$	$0.783^{\dagger}_{rBO}$	$0.550^{\dagger}_{rBO}$	$0.768^{\dagger}_{BO}$
	DDiv	D-ReDDE.top	$0.711^{\dagger}_{B}$	$0.694^{\dagger}_{B}$	$0.725^{\dagger}_{B}$	$0.507^{\dagger}_{BO}$	$0.726^{\dagger}_{B}$
		D-CRCS	$0.778^{\dagger}_{rBO}$	$0.762^{\dagger}_{rBO}$	$0.791^{\dagger}_{rBO}$	$0.557^{\dagger}_{rBO}$	$0.782^{\dagger}_{rBO}$
	SDiv	S-ReDDE.top	$0.775^{\dagger}_{rBO}$	$0.757^{\dagger}_{rBO}$	$0.789^{\dagger}_{rBO}$	$0.567^{\dagger}_{rcCBO}$	$0.778^{\dagger}_{rCBO}$
		S-CRCS	$0.786^{\dagger}_{rBO}$	$0.769^{\dagger}_{rBO}$	$0.800^{\dagger}_{rBO}$	$0.544^{\dagger}_{rBO}$	$0.802^{\dagger}_{rRBO}$
		S-BigDoc	0.615	0.594	0.632	0.415	0.621
		S-CORI	0.653	0.634	0.669	0.396	0.676
	LR-DivS		$\mathbf{0.873}^{\dagger\ddagger}_{rcRCBO}$	$\mathbf{0.853}^{\dagger\ddagger}_{rcRCBO}$	$\mathbf{0.889}^{\dagger\ddagger}_{rcRCBO}$	$\mathbf{0.705}^{\dagger\ddagger}_{rcRCBO}$	$\mathbf{0.856}^{\dagger\ddagger}_{rcRCBO}$
Suggestions	Baseline	ReDDE.top	0.614	0.600	0.626	0.463	0.629
		CRCS	$0.768^{\dagger}_{BO}$	$0.751^{\dagger}_{BO}$	$0.782^{\dagger}_{BO}$	$0.555^{\dagger}_{BO}$	$0.768^{\dagger}_{BO}$
	DDiv	D-ReDDE.top	$0.724^{\dagger}_{B}$	$0.705^{\dagger}_{B}$	$0.739^{\dagger}_{B}$	$0.525^{\dagger}_{BO}$	$0.735^{\dagger}_{B}$
		D-CRCS	$0.781^{\dagger}_{rBO}$	$0.765^{\dagger}_{rBO}$	$0.794^{\dagger}_{rBO}$	$0.558^{\dagger}_{rBO}$	$0.787^{\dagger}_{rBO}$
	SDiv	S-ReDDE.top	$0.799^{\dagger}_{rBO}$	$0.779^{\dagger}_{rBO}$	$0.814^{\dagger}_{rBO}$	$0.590^{\dagger\ddagger}_{rcCBO}$	$0.800^{\dagger}_{rBO}$
		S-CRCS	$0.787^{\dagger}_{rBO}$	$0.771^{\dagger}_{rBO}$	$0.801^{\dagger}_{rBO}$	$0.550^{\dagger}_{rBO}$	$0.803^{\dagger}_{rBO}$
		S-BigDoc	0.614	0.593	0.632	0.411	0.624
		S-CORI	0.647	0.629	0.662	0.396	0.672
	LR-DivS		$\mathbf{0.821}^{\dagger\ddagger}_{rCBO}$	$\mathbf{0.805}^{\dagger\ddagger}_{rCBO}$	$\mathbf{0.834}^{\dagger\ddagger}_{rBO}$	$\mathbf{0.629}^{\dagger\ddagger}_{rcRCBO}$	$\mathbf{0.827}^{\dagger\ddagger}_{rBO}$

Table 5: Diversification results on document retrieval on all queries. The symbols $\dagger, \ddagger, r, c, R, C$ indicate statistical significance under paired t-test to ReDDE.top baseline, CRCS baseline, D-ReDDE.top, D-CRCS, S-ReDDE.top, and S-CRCS respectively ($p < 0.05$).

			ERR	α-nDCG	NRBP	P-IA	S-Recall
Given Subtopics	Baseline	ReDDE.top	0.246	0.276	0.226	0.111	0.409
		CRCS	$0.380^{\dagger}_{rR}$	$0.406^{\dagger}_{r}$	$\mathbf{0.365}^{\dagger}_{rR}$	$\mathbf{0.152}_{rR}$	$0.530^{\dagger}_{r}$
	DDiv	D-ReDDE.top	$0.319^{\dagger}$	$0.345^{\dagger}$	$0.304^{\dagger}$	$0.123^{\dagger}$	$0.477^{\dagger}$
		D-CRCS	$\mathbf{0.383}^{\dagger}_{rR}$	$\mathbf{0.410}^{\dagger}_{rR}$	$\mathbf{0.365}^{\dagger}_{rR}$	$0.151^{\dagger}_{rR}$	$\mathbf{0.535}^{\dagger}_{r}$
	SDiv	S-ReDDE.top	$0.357^{\dagger}_{r}$	$0.387^{\dagger}_{r}$	$0.340^{\dagger}_{r}$	$0.141^{\dagger}_{r}$	$0.526^{\dagger}_{r}$
		S-CRCS	$0.365^{\dagger}_{r}$	$0.395^{\dagger}_{r}$	$0.346^{\dagger}_{r}$	$0.142^{\dagger}_{r}$	$0.534^{\dagger}_{r}$
Suggestions	Baseline	ReDDE.top	0.249	0.277	0.230	0.114	0.400
		CRCS	$0.339^{\dagger}_{r}$	$0.378^{\dagger}_{r}$	$0.315^{\dagger}$	$\mathbf{0.157}^{\dagger}_{rR}$	$\mathbf{0.535}^{\dagger}_{r}$
	DDiv	D-ReDDE.top	$0.304^{\dagger}$	$0.333^{\dagger}$	$0.286^{\dagger}$	$0.129^{\dagger}$	$0.469^{\dagger}$
		D-CRCS	$\mathbf{0.356}^{\dagger}_{rR}$	$\mathbf{0.390}^{\dagger}_{rR}$	$\mathbf{0.335}^{\dagger\ddagger}_{rR}$	$\mathbf{0.157}^{\dagger}_{rR}$	$0.530^{\dagger}_{rR}$
	SDiv	S-ReDDE.top	$0.323^{\dagger}$	$0.359^{\dagger}_{r}$	$0.300^{\dagger}$	$0.144^{\dagger}_{r}$	$0.512^{\dagger}_{r}$
		S-CRCS	$0.342^{\dagger}_{r}$	$0.376^{\dagger}_{r}$	$0.320^{\dagger}$	$0.147^{\dagger}_{r}$	$0.530^{\dagger}_{r}$

Table 6: Diversification results on document retrieval on test queries. The symbols $\dagger, \ddagger, r, c, R, C, B, O$ indicate statistical significance under paired t-test to ReDDE.top baseline, CRCS baseline, D-ReDDE.top, D-CRCS, S-ReDDE.top, S-CRCS, S-BigDoc, and S-CORI respectively ($p < 0.05$).

			ERR	α-nDCG	NRBP	P-IA	S-Recall
Given Subtopics	Baseline	ReDDE.top	0.304	0.345	0.278	0.151	0.509
		CRCS	$0.478^{\dagger}_{rBO}$	$0.512^{\dagger}_{rBO}$	$0.457^{\dagger}_{rBO}$	$0.217^{\dagger}_{rRBO}$	$0.668^{\dagger}_{BO}$
	DDiv	D-ReDDE.top	$0.422^{\dagger}$	$0.457^{\dagger}$	$0.400^{\dagger}$	$0.178^{\dagger}_{O}$	$0.622^{\dagger}_{B}$
		D-CRCS	$0.485^{\dagger}_{rBO}$	$0.523^{\dagger}_{rBO}$	$0.460^{\dagger}_{rBO}$	$0.219^{\dagger}_{rRBO}$	$0.677^{\dagger}_{rBO}$
	SDiv	S-ReDDE.top	$0.463^{\dagger}_{BO}$	$0.501^{\dagger}_{rBO}$	$0.438^{\dagger}_{BO}$	$0.200^{\dagger}_{rBO}$	$0.678^{\dagger}_{rBO}$
		S-CRCS	$0.473^{\dagger}_{rBO}$	$0.515^{\dagger}_{rBO}$	$0.446^{\dagger}_{BO}$	$0.206^{\dagger}_{rBO}$	$0.687^{\dagger}_{rBO}$
		S-BigDoc	0.370	0.390	0.359	0.151	0.519
		S-CORI	0.368	0.393	0.354	0.129	0.545
	LR-DivS		$\mathbf{0.515}^{\dagger}_{rRCBO}$	$\mathbf{0.555}^{\dagger}_{rRCBO}$	$\mathbf{0.489}^{\dagger}_{rRCBO}$	$\mathbf{0.243}^{\dagger\ddagger}_{rcRCBO}$	$\mathbf{0.722}^{\dagger\ddagger}_{rRBO}$
Suggestions	Baseline	ReDDE.top	0.337	0.373	0.311	0.167	0.518
		CRCS	$0.469^{\dagger}_{rBO}$	$0.509^{\dagger}_{rBO}$	$0.445^{\dagger}_{BO}$	$0.235^{\dagger}_{rBO}$	$0.671^{\dagger}_{BO}$
	DDiv	D-ReDDE.top	$0.434^{\dagger}_{B}$	$0.471^{\dagger}_{BO}$	$0.410^{\dagger}_{B}$	$0.202^{\dagger}_{O}$	$0.637^{\dagger}_{B}$
		D-CRCS	$0.487^{\dagger}_{rBO}$	$0.524^{\dagger}_{rBO}$	$0.463^{\dagger}_{rBO}$	$0.233^{\dagger}_{rBO}$	$0.675^{\dagger}_{BO}$
	SDiv	S-ReDDE.top	$0.462^{\dagger}_{BO}$	$0.504^{\dagger}_{rBO}$	$0.435^{\dagger}_{BO}$	$0.224^{\dagger}_{BO}$	$0.674^{\dagger}_{BO}$
		S-CRCS	$0.472^{\dagger}_{BO}$	$0.512^{\dagger}_{rBO}$	$0.448^{\dagger}_{BO}$	$0.219^{\dagger}_{BO}$	$0.675^{\dagger}_{BO}$
		S-BigDoc	0.347	0.374	0.330	0.157	0.511
		S-CORI	0.370	0.400	0.349	0.150	0.560
	LR-DivS		$\mathbf{0.509}^{\dagger\ddagger}_{rRCBO}$	$\mathbf{0.543}^{\dagger\ddagger}_{rRCBO}$	$\mathbf{0.487}^{\dagger\ddagger}_{rRBO}$	$\mathbf{0.249}^{\dagger}_{rBO}$	$\mathbf{0.690}^{\dagger}_{rBO}$

posed for resource selection in federated search, but limited attention has been paid to result novelty and diversification, which affects the effectiveness of existing algorithms. As far as we know, this paper proposes the first piece of research for incorporating search result diversification in resource selection for federated search.

A family of new evaluation metrics is first proposed for measuring search result diversification in resource selection, which combines some popular diversification metrics in ad hoc search with the recall-based evaluation metric in resource selection. Two general approaches are then proposed for diversification in selecting relevant sources. The first approach is based on sample documents, which ranks sample documents with respect to result diversification, and then utilizes the ReDDE framework for ranking the sources. The second approach is based on source-level estimation, which directly ranks each information source as a whole for result diversification. Furthermore, a learning based classification approach is proposed to combine multiple resource selection algorithms for more accurate diversification results.

An intensive set of empirical studies has been conducted to evaluate the proposed research on the Clueweb federated search dataset. Both the approach based on sample documents and on source-level estimation can outperform traditional resource selection algorithms in result diversification in both source-level for resource selection and in the document-level for federated document retrieval. Moreover, the learning based approach, which combines outputs

of multiple resource selection algorithms for result diversification, has been shown to generate the best results when some training data is available.

There are several possible directions to pursue in the future. The learning based method in this paper utilizes a simple model for combining outputs of multiple algorithms for result diversification, while a more sophisticated learning method may be more effective. Furthermore, it is an interesting topic to design new result merging algorithms with the focus on result diversification.

9. ACKNOWLEDGMENTS

This work is partially supported by NSF research grants IIS-0746830, CNS- 1012208 and IIS-1017837. This work is also partially supported by the Vietnam Education Foundation, the Center for Science of Information (CSoI), an NSF Science and Technology Center, under grant agreement CCF-0939370, and a travel grant from the ACM Special Interest Group on Information Retrieval. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsors.

10. REFERENCES

[1] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong. Diversifying search results. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining* pages 5–14, 2009.

[2] J. Arguello, J. Callan, and F. Diaz. Classification-based resource selection. *CIKM'09*, pages 1277–1286, 2009.

[3] M. Baillie, M. Carman, and F. Crestani. A multi-collection latent topic model for federated search. *Information Retrieval*, 14(4):390–412, 2011.

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

[5] J. Carbonell and J. Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *SIGIR'98*, pages 335–336, 1998.

[6] B. Carterette and P. Chandar. Probabilistic models of ranking novel documents for faceted topic retrieval. In *CIKM'09*, pages 1287–1296, 2009.

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

[8] H. Chen and D. Karger. Less is more: probabilistic models for retrieving fewer relevant documents. In *SIGIR'06*, pages 429–436, 2006.

[9] C. Clarke, N. Craswell, and I. Soboroff. Overview of the TREC 2009 Web Track. *TREC*, pages 1–9, Jan. 2009.

[10] C. Clarke, N. Craswell, I. Soboroff, and G. V. Cormack. Overview of the TREC 2010 Web Track. *TREC*, pages 1–9, Jan. 2010.

[11] C. Clarke, N. Craswell, I. Soboroff, and E. Voorhees. Overview of the TREC 2011 Web Track. pages 1–9, Jan. 2011.

[12] C. Clarke, M. Kolla, and O. Vechtomova. An effectiveness measure for ambiguous and underspecified queries. *Advances in Information Retrieval Theory*, pages 188–199, 2009.

[13] C. L. A. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, and I. MacKinnon. Novelty and diversity in information retrieval evaluation. In *SIGIR'08*, pages 659–666, 2008.

[14] N. Craswell. *Methods for Distributed Information Retrieval.* PhD thesis, The Australian National University, 2000.

[15] F. Crestani and I. Markov. Distributed Information Retrieval and Applications. In *Proceedings of ECIR*, Jan. 2013.

[16] V. Dang and W. B. Croft. Diversity by proportionality: an election-based approach to search result diversification. In *SIGIR'12*, pages 65–74. ACM, 2012.

[17] N. Fuhr. Resource Discovery in Distributed Digital Libraries. In *In Digital Libraries '99: Advanced Methods and Technologies, Digital Collections*, 1999.

[18] A. Genkin, D. D. Lewis, and D. Madigan. Large-scale Bayesian logistic regression for text categorization. *Technometrics*, 49(3):291–304, 2007.

[19] J. He, V. Hollink, and A. de Vries. Combining implicit and explicit topic representations for result diversification. In *SIGIR'12*, pages 851–860. ACM, 2012.

[20] D. Hong, L. Si, P. Bracke, M. Witt, and T. Juchcinski. A joint probabilistic classification model for resource selection. *SIGIR'10*, pages 98–105, 2010.

[21] A. Kulkarni and J. Callan. Document allocation policies for selective searching of distributed indexes. *CIKM'10*, pages 449–458, 2010.

[22] I. Markov, L. Azzopardi, and F. Crestani. Reducing the Uncertainty in Resource Selection. In *Proceedings of ECIR*, 2013.

[23] D. Metzler and W. B. Croft. Combining the language model and inference network approaches to retrieval. *Information Processing and Management*, 40(5):735–750, 2004.

[24] D. Nguyen, T. Demeester, D. Trieschnigg, and D. Hiemstra. Federated Search in the Wild. In *CIKM '12*, pages 1874–1878, 2012.

[25] R. L. Santos, C. Macdonald, and I. Ounis. Aggregated search result diversification. *Advances in Information Retrieval Theory*, pages 250–261, 2011.

[26] R. L. T. Santos, C. Macdonald, and I. Ounis. Exploiting query reformulations for web search result diversification. In *Proceedings of the 19th international conference on World wide web*, pages 881–890. ACM, 2010.

[27] M. Shokouhi. Central-rank-based collection selection in uncooperative distributed information retrieval. *Advances in Information Retrieval*, 2007.

[28] M. Shokouhi and L. Si. Federated Search. 2011.

[29] M. Shokouhi and J. Zobel. Federated Text Retrieval From Uncooperative Overlapped Collections. *SIGIR'07*, pages 789–790, 2007.

[30] M. Shokouhi and J. Zobel. Robust result merging using sample-based score estimates. *ACM Transactions on Information Systems (TOIS)*, 27(3):1–29, 2009.

[31] L. Si and J. Callan. A semisupervised learning method to merge search engine results. *ACM Transactions on Information Systems (TOIS)*, 21(4):457–491, 2003.

[32] L. Si and J. Callan. Relevant document distribution estimation method for resource selection. *SIGIR'03*, pages 298–305, 2003.

[33] P. Thomas and M. Shokouhi. Sushi: Scoring scaled samples for server selection. In *SIGIR'09*, pages 419–426. ACM, 2009.

[34] D. Vallet and P. Castells. Personalized diversification of search results. In *SIGIR'12*, pages 841–850. ACM, 2012.

[35] S. Vargas, P. Castells, and D. Vallet. Explicit relevance models in intent-oriented information retrieval diversification. In *SIGIR'12*, pages 75–84. ACM, 2012.

[36] J. Xu and W. B. Croft. Cluster-based language models for distributed retrieval. In *SIGIR'99*, pages 254–261, 1999.

[37] B. Yuwono and D. L. Lee. Server ranking for distributed text retrieval systems on the internet. In *Proceedings of the Fifth International Conference on Database Systems for Advanced Applications (DASFAA)*, pages 41–50, 1997.

[38] C. X. Zhai, W. Cohen, and J. Lafferty. Beyond independent relevance: methods and evaluation metrics for subtopic retrieval. In *SIGIR'03*, pages 10–17, 2003.

[39] K. Zhou, R. Cummins, M. Lalmas, and J. M. Jose. Evaluating aggregated search pages. In *SIGIR'12*, pages 115–124, 2012.